CHAPTER 1

Data presentation and collection

Objectives

When you have read this chapter you should be able to:

- construct tally charts and assemble data in the form of frequency tables;
- know the difference between discrete data and continuous data;
- present a visual display of frequency tables in the form of equal and unequal width histograms;
- use the histogram as a basis for constructing frequency polygons and frequency curves;
- develop and interpret cumulative frequency tables and their associated diagrams;
- present and use various visual approaches involving bar charts, pie charts, Lorenz curves, etc., which are widely used as a means of communicating patterns and trends in data.

Introduction

Statistics, as a subject, is a study of the various techniques of giving meaning to untreated or *raw* data. In this chapter we consider some simple, visual techniques which will be helpful in identifying patterns and trends in data. These will often involve drawing diagrams such as histograms, frequency curves, bar graphs, pie charts, Lorenz curves, etc. However we start by considering some simple procedures of data collection involving tally charts and the presentation of the data collected in a more 'grouped' form using class intervals and frequency tables. Ideas which involve selecting *samples* from all the data available and interpreting the results from sample data are considered in Chapter 7.

Answers to the 'Self-check questions', responses to each 'Pause for thought' and answers to the 'Review questions' can be found at the end of the book (pp. 320–458).

1.1 Frequency distribution

Frequency is the number of times a certain event has happened, and is often found by means of a *tally chart*. For example, if you were conducting a survey of defective items produced by machines of different ages (in months) in an assembly workshop, you might record your findings as shown in the tally chart of Figure 1.1.

Each time a defective item is identified as coming from a machine of a particular age, a *vertical* line is drawn, and should a fifth such line be needed a *diagonal* line is drawn through the previous four vertical lines. At the end of the data collection period the total number of tally marks are added for each age of machinery and the total placed in the end column marked F for frequency. The total frequency (80) is found by adding these individual (class) frequencies together.

figure 1.1
Tally chart for defective items

Age in months	*Tally*	*Frequency (F)*
0 and under 5	II	2
5 and under 10	III	3
10 and under 15	𝍸 II	7
15 and under 20	𝍸 IIII	9
20 and under 25	𝍸 𝍸 𝍸 𝍸	20
25 and under 30	𝍸 𝍸 𝍸 II	17
30 and under 35	𝍸 𝍸 𝍸 𝍸 II	22
		—
Total		80

1.2 Frequency tables

Figure 1.1 is essentially a **frequency table,** and it will be useful at the outset to define a number of terms often encountered when discussing such tables.

- *Class intervals*: these are the groups into which we organise the data. There are seven class intervals in Figure 1.1, the first being '0 and under 5' and so on.
- *Class boundaries*: these are the values at which the different class intervals meet. For example the second class interval has a lower class boundary (LCB) of 5 and an upper class boundary (UCB) of 10.
- *Class limits*: these are the lower and upper values of the class interval *as shown in the frequency table.* In Figure 1.1 the class limits are the same as the class boundaries. However this may not always be the case. Suppose a frequency table contains the following entries.

Marks	*Frequency*
0 to 9	5
10 to 19	6
20 to 29	14
30 to 39	18

The second class interval has *class limits* of 10 and 19 as shown in the frequency table, but *class boundaries* of 9.5 and 19.5. Where there is a gap in the values between class limits, the mid-point of the gap will be the value at which the class intervals meet, i.e. the class boundary. For example a mark below 9.5 will go into the first class interval, whereas a mark of 9.5 or above will go into the second class interval.

- *Class width (or size)*: this is the difference between the upper and lower class boundaries for each class interval. As we can see from Figure 1.1 all the class intervals have a class width (or size) of 5.
- *Class frequency*: this is the number of observations found to occur in a particular class interval.

A frequency table is useful in giving an immediate impression of how the raw data are distributed. We can see from Figure 1.1 that, although the relationship is not precise, the older the machine the greater the number of defective items that were found to occur.

The following steps may help when you try to construct your own frequency table from raw data.

Step 1 Determine the range of the data:

range = maximum value – minimum value

Step 2 Determine the number of class intervals to be shown in the table. This is largely a matter of opinion – too many class intervals will confuse the data whilst too few class intervals will mean that you lose much of the information. Most published frequency tables use between five and ten class intervals.

Step 3 Determine the width (size) of class intervals.
As a general rule the width of each class interval should be approximately equal to the range of the data divided by the number of class intervals.

Sometimes Steps 2 and 3 are best reversed, as when you have already decided on an appropriate width (size) for your class interval.

Step 4 Determine the first lower class limit.

Step 5 Construct the frequency table using a tally chart.

WORKED EXAMPLE **1.1**

The percentage marks scored by 58 candidates seeking promotion in a personnel selection test were recorded as follows:

37	49	58	59	56	79
62	82	53	58	34	45
40	43	44	50	42	61
54	30	49	54	76	47
64	53	64	54	60	39
49	44	47	44	25	38
55	57	54	55	59	40
31	41	53	47	58	55
59	64	56	42	38	37
33	33	47	50		

Construct a frequency table of the data, using class intervals with a class width (size) of ten percentage marks.

Solution

Step 1 Determine the range of the data:

maximum = 82; minimum = 25; range = 57

Steps 2 and 3 Determine the width (size) and number of the class intervals. Width of class intervals = 10 (e.g. 20.5 to 30.5). Therefore seven class intervals will be required to cover scores 21 to 90 (per cent) should we decide to set the first lower class limit at 21, as in Step 4.

Step 4 First lower class limit = 21

Step 5 Construct the frequency table

figure 1.2
Tally chart for percentage scores

Class	Tally	Frequency
21–30	II	2
31–40	𝍸 𝍸 I	11
41–50	𝍸 𝍸 𝍸 II	17
51–60	𝍸 𝍸 𝍸 𝍸	20
61–70	𝍸	5
71–80	II	2
81–90	I	1

1.3 Discrete or continuous data

Data are either 'discrete' or 'continuous'.

Discrete data

These are data which can only take on a *limited number* of different values. For example:

- the number of houses built
- the number of goals scored
- the number of marks scored in a test

All these examples can only take on a limited or finite number of *integer* (whole number) values.

Continuous data

These are data which can take on an *infinite number* of different values (at least in theory):

- your height
- your weight
- your age
- the time taken to do a job

All these examples can take on *any* value, subject only to the minimum unit of measurement available. These values need *not* be integers; example, 0.001 cm, 1.24 grams, 4.03 seconds, and so on. Continuous data are often 'rounded off' to a

DID YOU KNOW?

There has been a huge increase in the volume of data a firm must process. The OECD has estimated that it has grown sevenfold in the past 20 years. Despite the growth of IT, some 92 million paper documents involving data of one kind or another are produced each year.

certain number of decimal places even though still smaller measurements might theoretically be possible.

PAUSE FOR THOUGHT 1.1 *Can you give two more examples of discrete data and two more examples of continuous data?*

1.4 Histograms

The **histogram** is a visual means which is often used to display data which have been arranged in the form of a frequency table. The histogram can be used for displaying both *discrete* and *continuous* data. The histogram has a number of properties:

- it is a set of rectangles, each of which represents the frequency of a particular class interval;
- each rectangle is constructed so that its *area* is in proportion to the frequency of the class interval it represents;
- the base of each rectangle represents the width (or size) of the class interval and the height of each rectangle represents the frequency of the class interval;
- when all the class intervals have the same width (or size), then the vertical axis which represents the height of each rectangle is the class frequency;
- when the class intervals are of *unequal width* (or size), then the height of each rectangle must be adjusted where it differs from the 'standard' class width, i.e. from the class width of the majority of class intervals. For example if the width of a particular class interval *doubles* (and with it the base of the rectangle), then we must *halve* the height of the rectangle used to represent its class frequency. We must do this to keep the *areas* of the rectangles proportional to the class frequencies. The vertical axis is no longer the frequency of the class interval but the *frequency density*, as we shall see below.

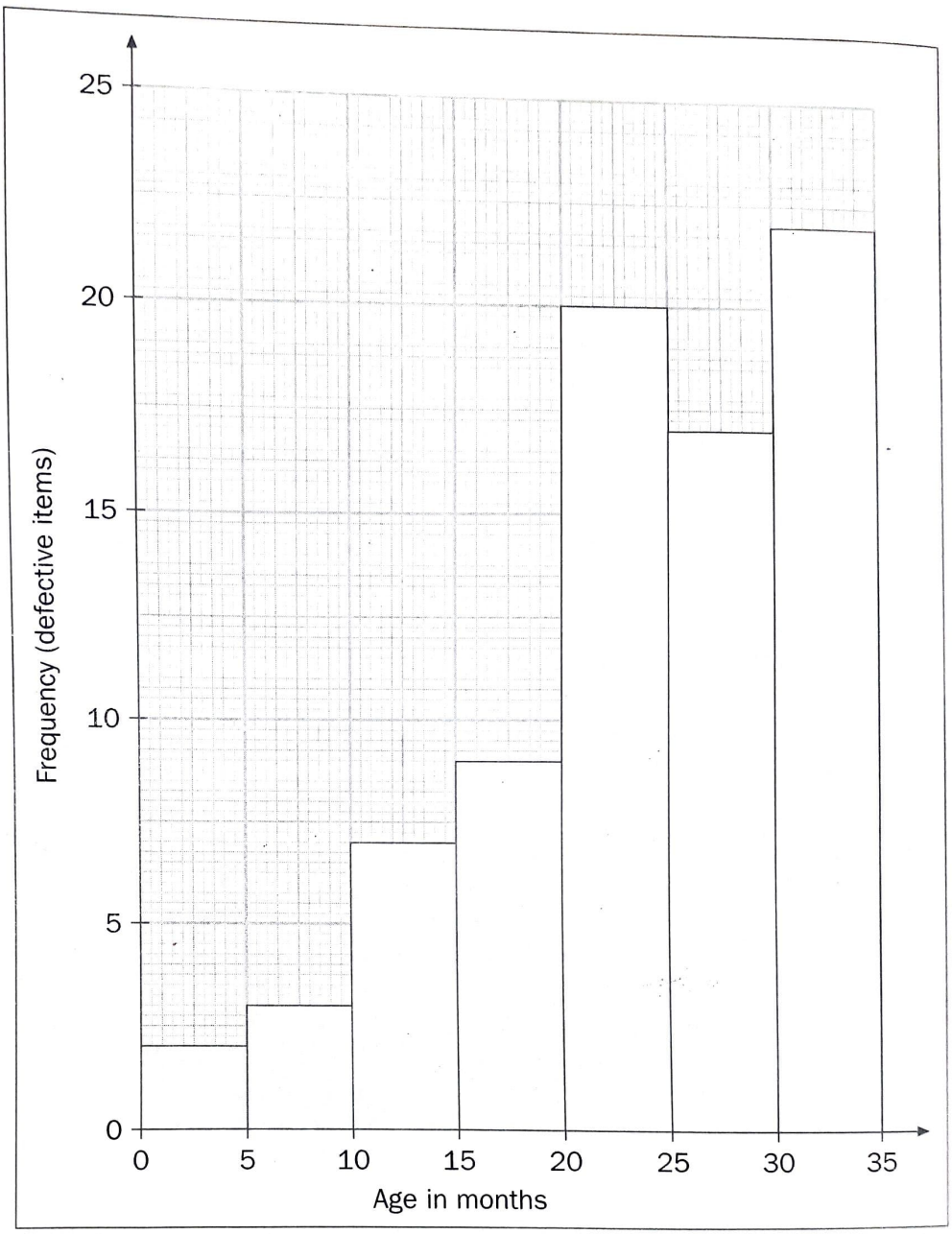
Equal-width histograms

Table 1.1 represents the data from our earlier example of defective items produced by machines of varying age.

table 1.1
Frequency table for defective items by age of machine

Age in months	*Frequency*
0–5	2
5–10	3
10–15	7
15–20	9
20–25	20
25–30	17
30–35	22
	80

figure 1.3
Equal-width histogram



As can be seen, each class interval has an *equal* class width (or size) of five months. This is sometimes called the *standard class width*. The **equal-width histogram** which represents these data is shown in Figure 1.3.

Note that the rectangles are connected together, that the base on each rectangle represents the standard class width (five months) and is the same throughout, and that the height of each rectangle represents the class frequency. As a result the *areas* of the rectangles are in proportion to the respective class frequencies, and so we can see at a glance that the bulk of the frequency distribution is to the right of the histogram, in the rectangles 20–25, 25–30 and 30–35 months. We gain an immediate visual impression from the histogram that the frequency (of defective items) increases broadly in relation to the age of the machinery.

table 1.2
Frequency table for time taken by employees to complete a task

Time (nearest minute)	Frequency
0–1	42
2	35
3	30
4	21
5	15
6–7	15
8–9	6
10	1
	165

table 1.3
Frequency density for time taken by employees to complete a task

Time (nearest minute)	Frequency	Class width (minutes)	Frequency density
0–1	42	1.5	(42 ÷ 1.5) 28
2	35	1	35
3	30	1	30
4	21	1	21
5	15	1	15
6–7	15	2	(15 ÷ 2) 7.5
8–9	6	2	(6 ÷ 2) 3
10	1	1	1

Unequal-width histogram

Sometimes the data will be such that the class intervals are of different widths (or sizes). In the previous example of defective items the width of each class interval was five months. In the next example (Table 1.2) however we have a frequency table in which the widths of the various class intervals vary, being either 1 minute, 1.5 minutes or 2 minutes.

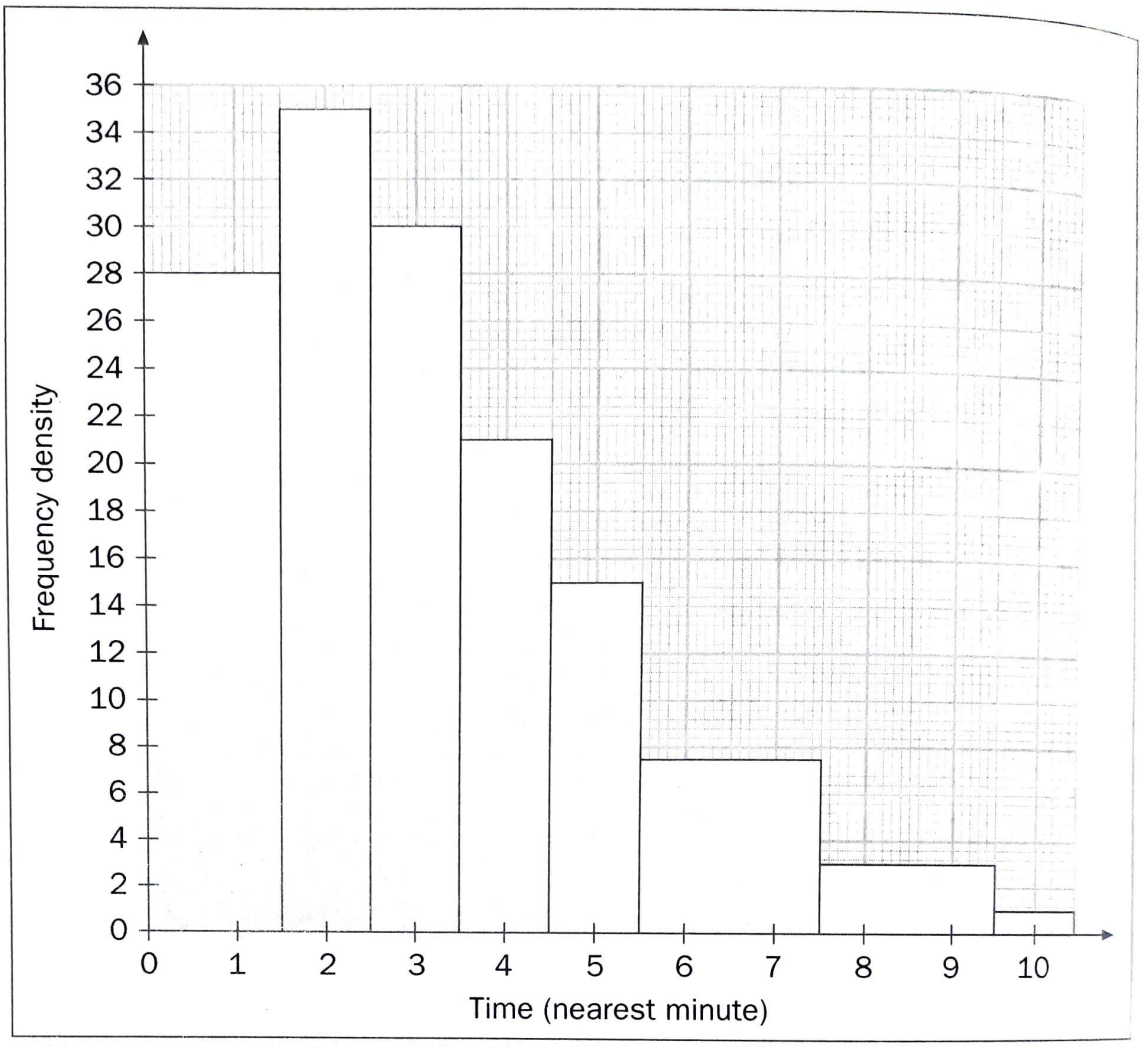
Remembering that the *areas* of the rectangles must be in proportion to the class frequencies, we must adjust the height of any rectangles in line with any variations in the base of that rectangle. So if the base (class width) doubles in length we must halve the height (class frequency) and so on. The value on the vertical axis is no longer frequency, but *frequency density*. These adjustments are shown in Table 1.3 and Figure 1.3.

Remembering that the *class width* is the difference between upper and lower *class boundaries*, we can see that the first class interval has a width of 1.5 minutes, and the sixth and seventh class intervals have class widths of 2 minutes, with all other intervals having a class width of 1 minute. The adjusted frequency densities from Table 1.3 are shown in the **unequal-width** histogram below (Figure 1.4).

The class width which appears most often in an unequal-width frequency table and histogram is often called the *standard class width*. In the above example the standard class width is 1 minute. We derive the **frequency density** by dividing the class frequency by the number of standard class widths in that class interval. So, for example, we divide the six employees in the penultimate class interval by two to give a frequency density of three.

This approach can be seen more clearly in the following worked example.

figure 1.4
Unequal-width histogram



WORKED EXAMPLE **1.2**

A company wishes to review its distribution operation and from its time sheet records it found that 144 vehicles were loaded in a 24 hour period. A frequency distribution table was prepared from the data as follows.

Time to load (minutes)	*Number of vehicles*
40–45	4
45–50	13
50–55	17
55–60	44
60–70	59
70–80	7

From the above data construct a histogram to represent the data.

Solution

Step 1 From the frequency table decide on a standard class width = 5 minutes.
Step 2 Enlarge the original table, adding columns for class width, number of standard class widths and height (frequency density).

Time	*Frequency*	*Class width*	*No. of standard class widths*	*Height (frequency density)*
40–45	4	5	1	4
45–50	13	5	1	13
50–55	17	5	1	17
55–60	44	5	1	44
60–70	59	10	2	(59 ÷ 2) 29.5
70–80	7	10	2	(7 ÷ 2) 3.5

Note that where the standard class width has doubled, as in the case of the last two class intervals, we must halve the height (frequency density) of the corresponding rectangle.

Step 3 Construct the histogram

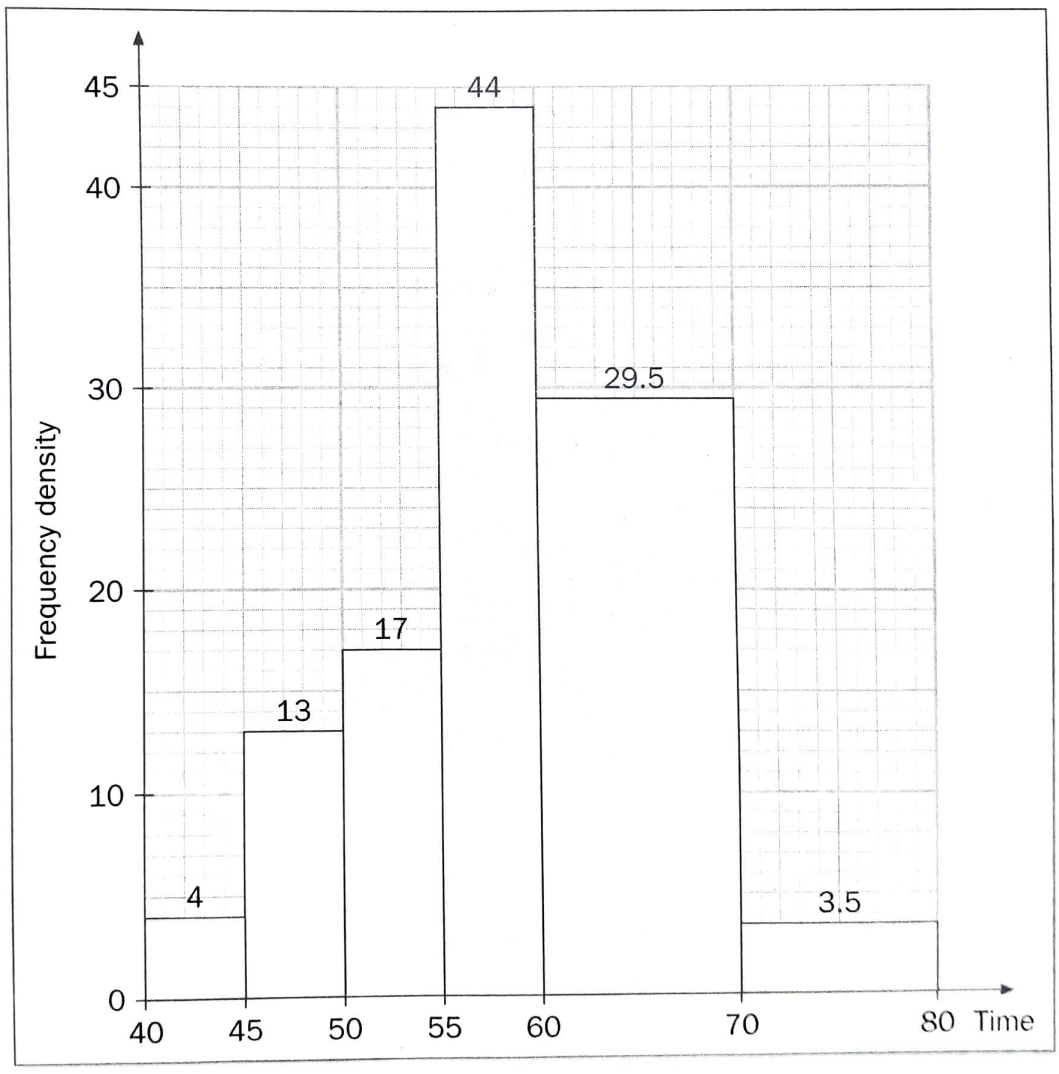


figure 1.5
Histogram showing number of vehicles loaded in a 24 hour period

1.5 Frequency polygon

A **frequency polygon** is a straight line diagram which represents exactly the same area (i.e. frequency) as shown in the histogram. The frequency polygon is derived from the histogram by connecting the mid-points of the tops of each

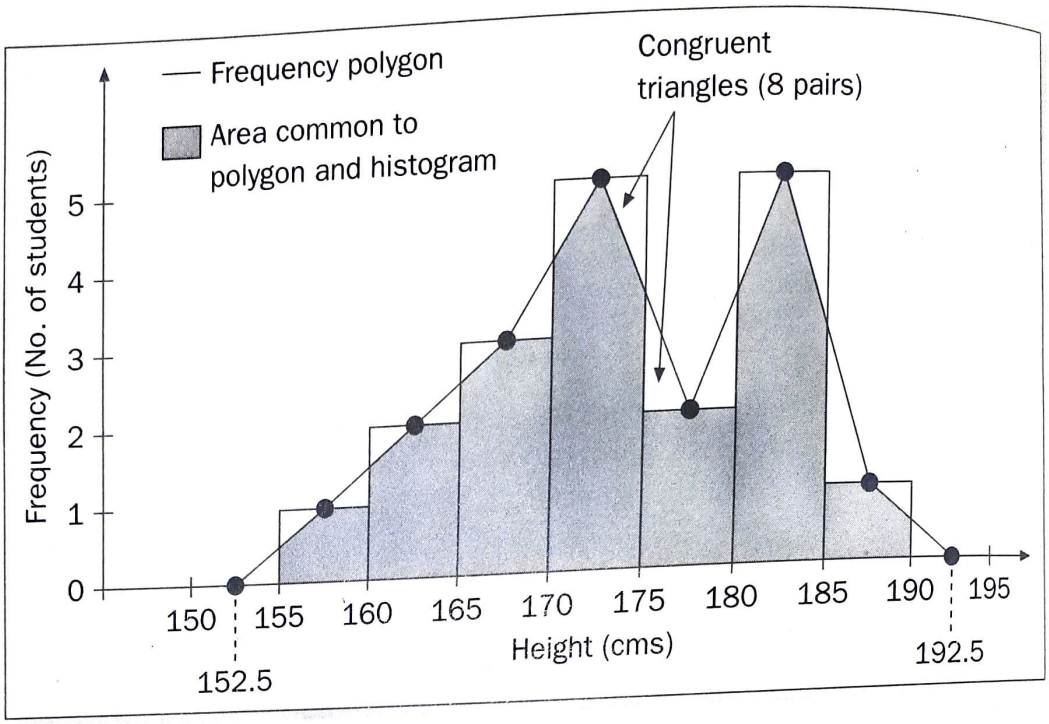


figure 1.6
Frequency polygon: area enclosed identical to histogram

adjacent rectangle by straight lines, as in Figure 1.6. If we then use straight lines to join the mid-points of the tops of the end two rectangles with the horizontal axis, as shown, then a key property of the frequency polygon should quickly become apparent: namely that the area under the histogram is identically equal to the area under the frequency polygon.

The shaded area is common to both polygon and histogram and there are eight respective pairs of congruent triangles (side–angle–side), giving identical areas contained by polygon and histogram.

1.6 Frequency curve

From the previous discussion we can therefore state that the frequency distribution represented by the histogram can equivalently be represented by the straight line polygon known as the **frequency polygon**. Further, if the points to be connected by straight lines were close together then, as an approximation, we could represent the frequency polygon by the **frequency curve** shown in Figure 1.7. Indeed the frequency curve is the most usual way of representing a frequency table visually. However, it is basically an alternative means of displaying visually the properties of a frequency table already captured by the histogram and frequency polygon.

The two peaks indicated by the frequency curve would represent a *bimodal* distribution, as we note in Chapter 2.

1.7 Cumulative frequency curves

Cumulative frequency is the 'running total' of the figures shown in the frequency column of a frequency table. As we shall see we can draw up cumulative 'more than' and cumulative 'less than' tables to represent data on income,

figure 1.7
Frequency curve as an approximation to the frequency polygon

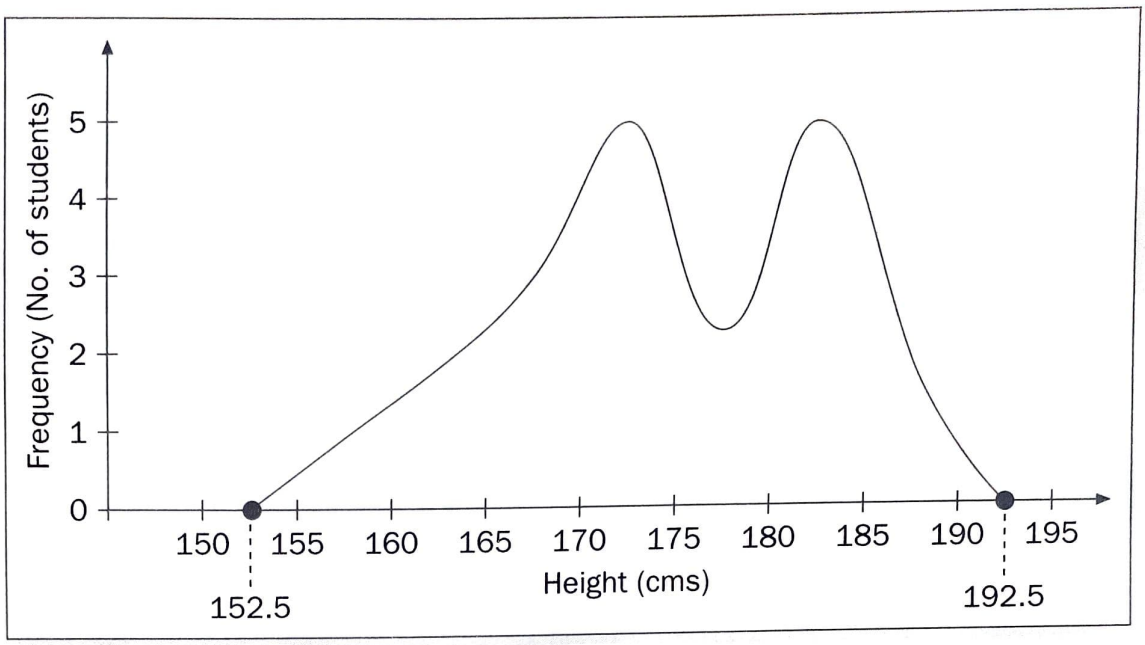


table 1.4
Frequency table for height of trees (m)

Height (m)	*0–1.9*	*2–2.4*	*2.5–2.9*	*3.0–3.4*	*3.5–4.5*
Frequency	15	43	30	10	2

table 1.5
Cumulative 'less than' frequency table for height of trees (m)

Height (m)	*Frequency*	*Height less than (m)*	*Cumulative frequency*
0–1.9	15	1.95	15
2–2.4	43	2.45	58
2.5–2.9	30	2.95	88
3.0–3.4	10	3.45	98
3.5–4.5	2	4.55	100

wealth and many other variables. We can then draw cumulative frequency *polygons* or *curves* to give a visual picture of these tables.

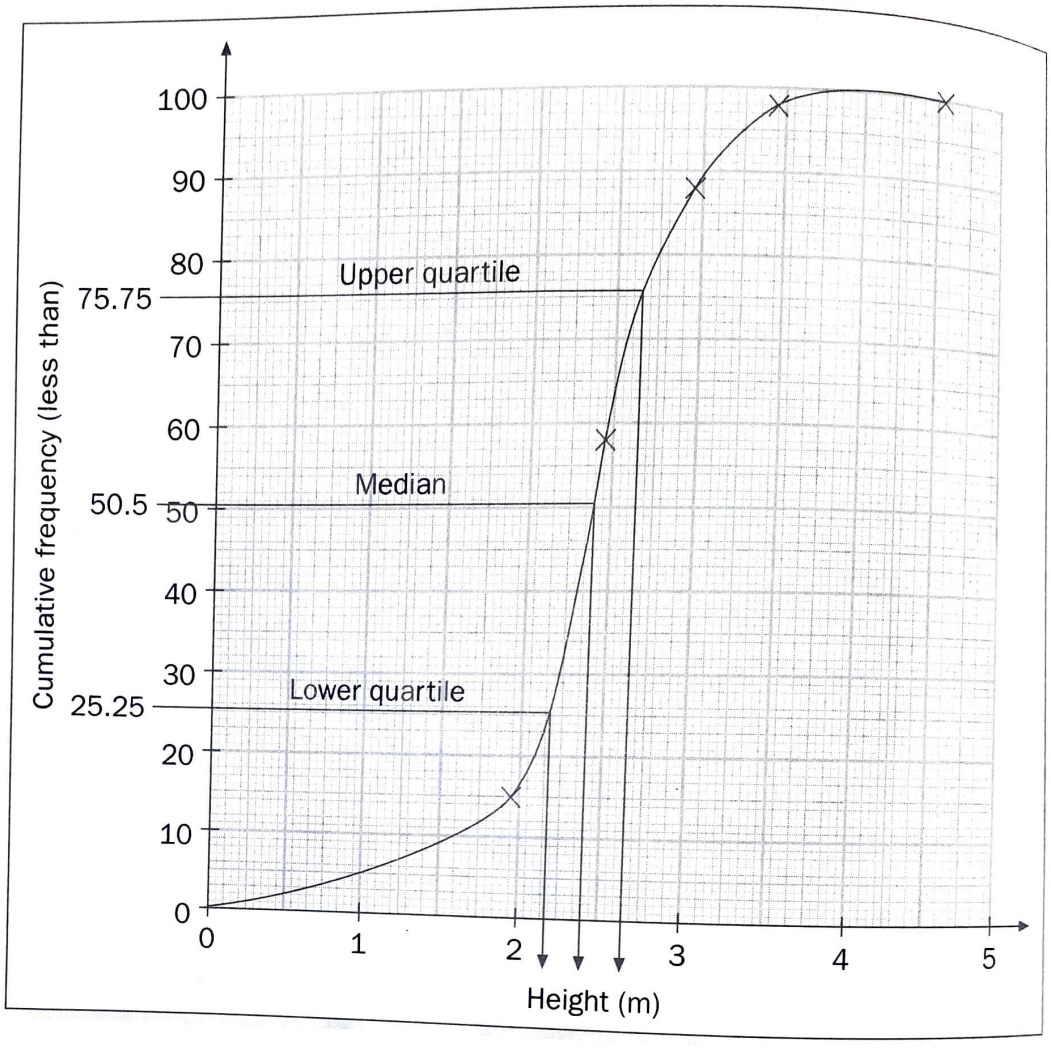
Table 1.4 illustrates the frequency distribution of the height of 100 trees in a plantation, each tree being measured to the nearest 0.1 m.

We can, for example, produce a *cumulative 'less than' frequency table* (Table 1.5). This keeps a running total on how many trees are 'less than' a particular height. As we saw with histograms, the width (size) of the class intervals (groups) are slightly wider than the table suggests. For example 2–2.4 is really $1.95 \leq$ height < 2.45, i.e. a class width of 0.5 (m).

We now draw a cumulative frequency *curve* from this table by plotting the cumulative frequency (on the vertical axis) against the 'less than' value (on the horizontal axis). Here, we join up the points we have plotted with a smooth curve. This frequency curve is also known as an '**ogive**'. The curve plotted in Figure 1.8 is therefore the cumulative 'less than' curve (or ogive).

You can take note of some useful information from such a curve. For example all but two of the trees (i.e. 98 out of a hundred) have a height less

figure 1.8
Cumulative 'less than' curve or ogive



than 3.45 m. On the other hand, only 15 out of a hundred trees are less than 1.95 m in height.

PAUSE FOR THOUGHT 1.2 *Try to construct a 'cumulative more than' frequency table using the data in Table 1.5 but using the lower class boundaries for each class interval.*

Percentiles

We can also use our cumulative 'less than' frequency curve to answer questions involving *percentiles*. A percentile is a particular value below (or above) which a given percentage of the distribution lies.

For example, 'what is the height of tree for which 25 per cent of the distribution will lie below that height?' This height is the **lower quartile** height (or 'first quartile'), shown as 2.15 m in Figure 1.8. Note that since the total frequency is 100, then 25 per cent is 25 trees in this case. However because of the *continuity correction* (see box below) we should strictly regard 25.25 trees as the lower quartile observation. Let us consider the continuity correction for first, second (median) and third quartiles respectively, before using Figure 1.8 to estimate those values.

Continuity correction

Strictly speaking we should make slight adjustments when finding the quartile values of a distribution:

- lower (first) quartile position $\frac{1}{4}(n+1)$
- median (second) quartile position $\frac{1}{2}(n+1)$
- upper (third) quartile position $\frac{3}{4}(n+1)$

where n is the number of observations.

It is perhaps helpful to explain the reason for this continuity correction by starting with the second quartile, the median.

Median

The **median** (second quartile) value is that value for which 50 per cent of the data are below it and 50 per cent above it. If we arrange the items of data in *order* (whether rising or falling in value), we need to find the value corresponding to the middle item. For example, with nine items, the median value will correspond to that of the fifth item arranged in order, i.e. the value 7 in the example below.

2 4 4 5 (7) 10 11 14 18

Note that the median position is given by $\frac{1}{2}(n+1) = \frac{1}{2}(9+1) = 5$th item arranged in order. However when the number of observations, n, is not an odd number but an even number, then the median value will lie mid-way between two observations. For example, with ten items as a result of adding the number 20 to the previous nine items, the median value will correspond to that of the 5.5th item arranged in order.

median position

2 4 4 5 7 ↕ 10 11 14 18 20

5.5th item

Notice that the median position is given by $\frac{1}{2}(n+1) = \frac{1}{2}(10+1) = 5.5$th item arranged in order.

We can therefore say that the median position is given by $\frac{1}{2}(n+1)$, where this expression is known as the continuity correction.

In the previous example with 100 trees, the median item in the cumulative frequency table is the $\frac{1}{2}(100+1) = 50.5$th item.

We can use our cumulative frequency curve (Figure 1.8) to find the median height of trees. We draw a horizontal line from 50.5 on the vertical (cumulative frequency) axis to the curve. We then draw a vertical line from the curve to the

horizontal axis and 'read off' the value, i.e. around 2.4 m. So half the number of trees are below 2.4 m in height, and half above 2.4 m in height.

We follow a similar procedure for finding the first and third quartile positions as we do for the second quartile (median). However instead of $\frac{1}{2}$ (i.e. $\frac{2}{4}$) we have $\frac{1}{4}$ and $\frac{3}{4}$ respectively.

PAUSE FOR THOUGHT 1.3 *At what value would you expect a cumulative 'more than' frequency curve to intersect a cumulative 'less than' frequency curve?*

Lower quartile

The **lower quartile** (first quartile) is that value for which a quarter (25 per cent) of the distribution lies *below* it. If there are n items in the distribution, we need to find the value corresponding to the $\frac{1}{4}(n + 1)$th item, when the items are arranged in order. In our 'tree' example, $n = 100$ so $\frac{1}{4}(n + 1) = \frac{101}{4} = 25.25$.

We can therefore use our cumulative frequency curve to estimate the value of the lower quartile. This time draw the horizontal line from 25.25 on the vertical (cumulative frequency) axis to the curve, and read down to around 2.15 metres on the horizontal axis. So a quarter of the trees are below 2.15 metres in height.

Upper quartile

The **upper quartile** (third quartile) is that value for which a quarter (25 per cent) of the distribution lies *above* it. (We therefore look three-quarters of the way up the distribution, arranged in order, where we find the upper quartile). If there are n items in the distribution, we need to find the value corresponding to the $\frac{3}{4}(n + 1)$th item, when the items are arranged in order. If our 'tree' example, $n = 100$ so $\frac{3}{4}(100 + 1) = \frac{303}{4} = 75.75$.

In Figure 1.8 we now read across from 75.75 on the vertical (cumulative frequency) axis to the curve, and then down to the horizontal axis. This gives a reading of around 2.65 m. So one-quarter of the trees are above 2.65 metres in height (i.e. three-quarters of trees are below 2.65 metres in height).

We can, in these ways, find the heights of trees at *any percentile* point of the distribution. Here we have used *quartiles* (25 per cents), but we could have used *quintiles* (20 per cents) or *deciles* (10 per cents).

Of course we can use the cumulative 'less than' distribution the other way round, this time starting on the horizontal axis and then moving to the vertical axis. For example, we can read off the number (here percentage) of trees having a height less than a specific value. For example 90 trees (90 per cent) have a height less than 3 m.

We return to the ideas of median and lower and upper quartiles in Chapter 2.

SELF-CHECK QUESTIONS

1.1 The frequency table below shows the number and value of transactions at a corner shop taking place between the hours of 17.00 and 22.00.

Value of transaction (£)	*Frequency*
0–3.99	8
4–7.99	18
8–11.99	22
12–15.99	24
16–19.99	13
20–23.99	12
24–27.99	2
28–31.99	1

Construct a histogram of the data.

1.2 Here is the frequency table of the weekly income of 45 people in a survey. Draw a histogram to represent these data (*note*: unequal-width histogram).

Weekly income (£)	*Frequency*
76–100	6
101–125	8
126–150	17
151–200	8
201–300	6

1.3 The following table shows the distribution of the ages of people attending a public concert.

Age range (yrs)	*0–19*	*20–39*	*40–59*	*60–79*	*80–99*
No. of people	8	26	110	128	56

Construct a cumulative frequency (less than) table and draw the corresponding curve on a graph. Use your curve to estimate:

(i) the median age attending the concert,
(ii) the lower and upper quartile ages,
(iii) the percentage of people over 65 attending the concert.

1.4 (a) Complete the cumulative 'more than' table below:

Height (m)	*Frequency*	*Height more than (m)*	*Cumulative frequency*
0–1.9	15	0.0	
2–2.4	43	1.95	
2.5–2.9	30	2.45	
3.0–3.4	10	2.95	
3.5–4.5	2	3.45	
		4.55	

(b) Draw the cumulative 'more than' curve on Figure 1.8.

(c) What do you notice about the point at which the cumulative 'less than' and 'more than' curves intersect?

Note: **Answers can be found on pp. 320–323.**

1.8 Bar charts

There are three main types of **bar chart**:

- simple bar chart
- component bar chart
- multiple bar chart

Simple bar chart

A **simple bar chart** is a set of non-joining bars of equal width whose height or length is proportional to the frequency it is representing. They can be drawn in a vertical or horizontal format and can show negative, as well as positive, values (see Figure 1.9).

Component bar chart

A **component or segmented bar chart** is useful to illustrate a breakdown in the figures, e.g. the total sales of XYZ Co. could be broken down into sales by product (see Figure 1.10(a)). The constituent parts of each bar are always stacked in the same order with the height of each representing the individual values or frequencies.

figure 1.9
Simple bar charts

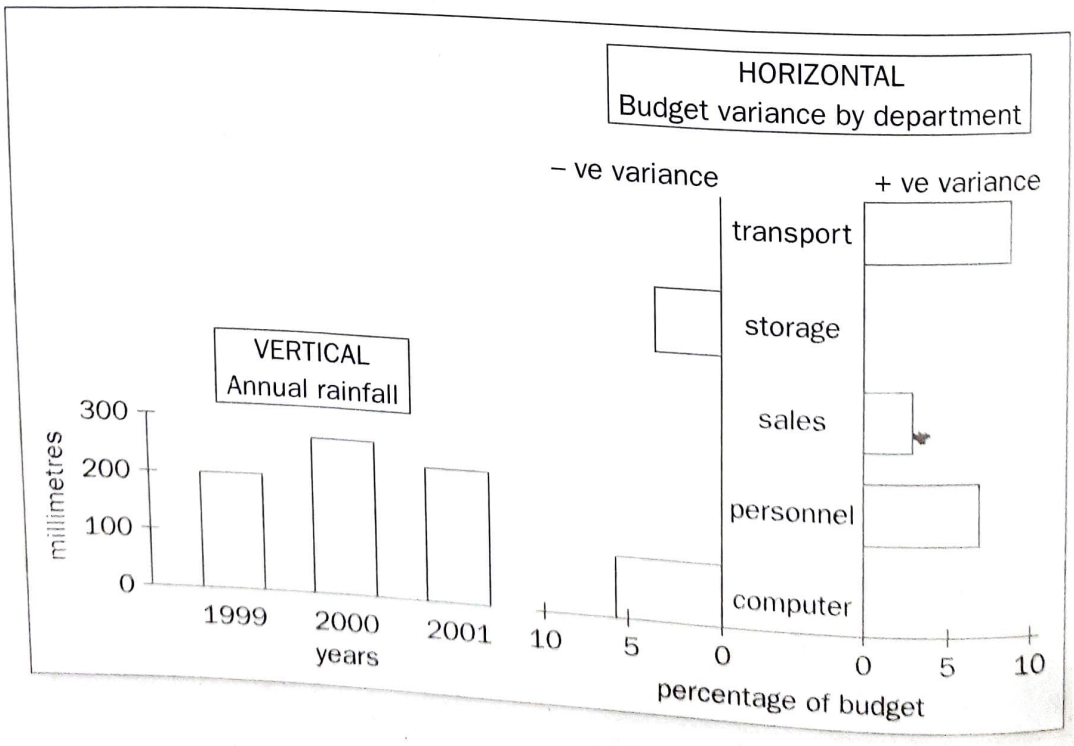
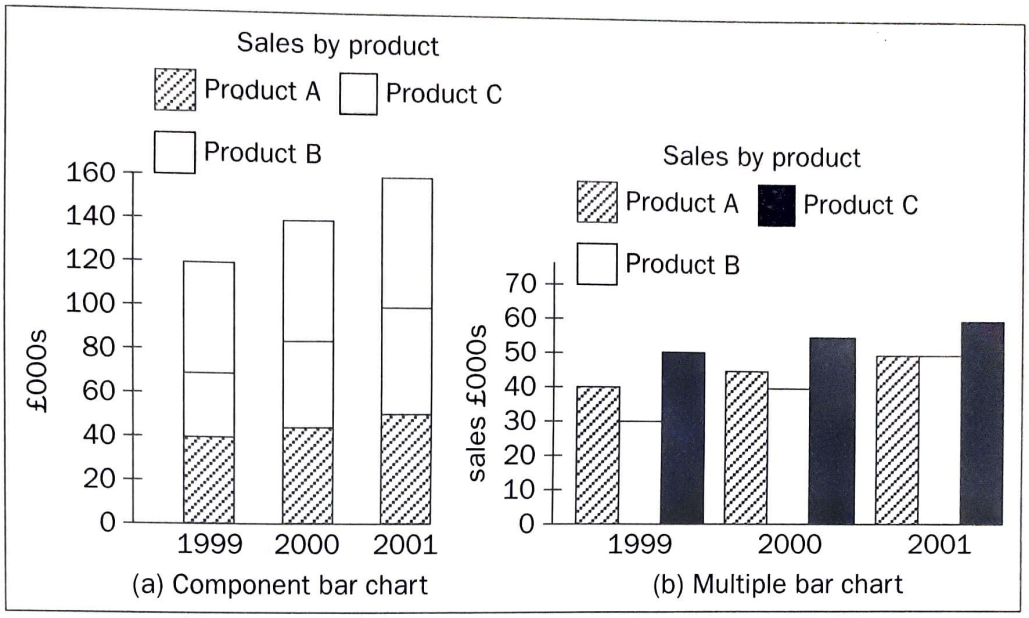


figure 1.10
Component and multiple bar charts



Multiple bar chart

A **multiple bar chart** uses a separate bar to represent each constituent part of the total. These bars are joined into a set for each class of data (see Figure 1.10(b)). The data in Figure 1.10(a) are represented in multiple bar chart form in Figure 1.10(b).

1.9 Pie chart

A **pie chart** involves a circle which is used to represent the *whole data*. Sectors of this circle ('pie') are then used to represent the relative frequencies of particular *components* of the data. The size of each sector is found by calculating the proportion of the whole data represented by the component and then multiplying by 360° to obtain the angle of the sector from the centre of the circle.

Pie charts give an immediate visual impression of the relative contribution of particular components to the overall total. They are widely used in company reports and accounts, as in highlighting the relative contribution of different product lines to overall profit or of different types of cost to overall cost. However they are only effective for displaying the contributions of a relatively small number of separate components.

WORKED EXAMPLE **1.3**

A survey of 12,000 television viewers on a particular evening found the following information as to the channels being watched at 8 p.m. on that evening. You are asked to construct a pie chart to represent these data.

table 1.6
TV viewing figures

Channel	*Number*
BBC1	3,000
BBC2	1,000
ITV3	5,500
ITV4	2,000
5	500

Solution

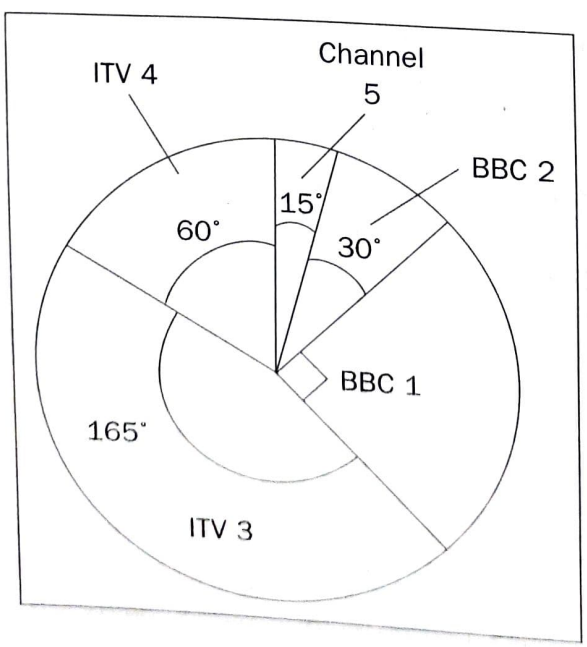
To find the angle of the circle ('pie') represented by each channel, we find the proportion of the whole data (12,000 viewers) represented by that channel and then multiply by 360°. This is done in Table 1.7.

table 1.7
Finding the sectors of the 'pie'

Channel	*Frequency*	*Angle*
BBC1	3,000	$\frac{3,000}{12,000} \times 360^\circ = 90^\circ$
BBC2	1,000	$\frac{1,000}{12,000} \times 360^\circ = 30^\circ$
ITV3	5,500	$\frac{5,500}{12,000} \times 360^\circ = 165^\circ$
ITV4	2,000	$\frac{2,000}{12,000} \times 360^\circ = 60^\circ$
Channel 5	500	$\frac{500}{12,000} \times 360^\circ = 15^\circ$

The pie chart is then drawn from this table using a protractor or angle measurer. The completed pie chart is shown in Figure 1.11. We can see at a glance how ITV3 dominated that evening's viewing figures, with channel 5 having the smallest share of viewers.

figure 1.11
Pie chart for viewing figures



In actually drawing the pie chart it is helpful to draw the smallest sectors first, so that any small inaccuracies in angles will be visually less important when you come to drawing the larger sectors.

ACTIVITY **1**

Greece is a major export market for your company. You sell a variety of products, many of which are age specific. You therefore require a demographic breakdown, by age, of the Greek population. Use the information in the following table and a spreadsheet with which you are familiar to plot the graphs mentioned.

Greece: population by age 1999	
	`000
0–9	1,160
10–19	1,425
20–29	1,544
30–39	1,428
40–49	1,263
50–59	1,305
60–69	1,081
70–79	663
80+	331

1 Construct an ogive (cumulative less than frequency curve) from the data in the table above (the first class [0 to 9] is placed in cell A4).
Remember that an ogive is plotted against the upper class boundary of the class interval therefore: put these upper class boundaries in column C (i.e. 9.5, 19.5, 29.5 . . .).
Calculate the cumulative less than frequency in column D as follows:
(a) in **cell D4** enter **1,160**,
(b) in **cell D5** construct a formula to calculate the cumulative less than frequency,
(c) copy the formula down to cell **D12**.
Plot a graph of the cumulative frequency, i.e. the ogive.
Use the ogive to estimate the population below 25 years of age.

2 Convert your results into a cumulative *percentage* less than ogive.

3 Construct a pie chart from the above data to show the relative importance of the age groups specified within the whole Greek population (total for 1999: 10,200,000).

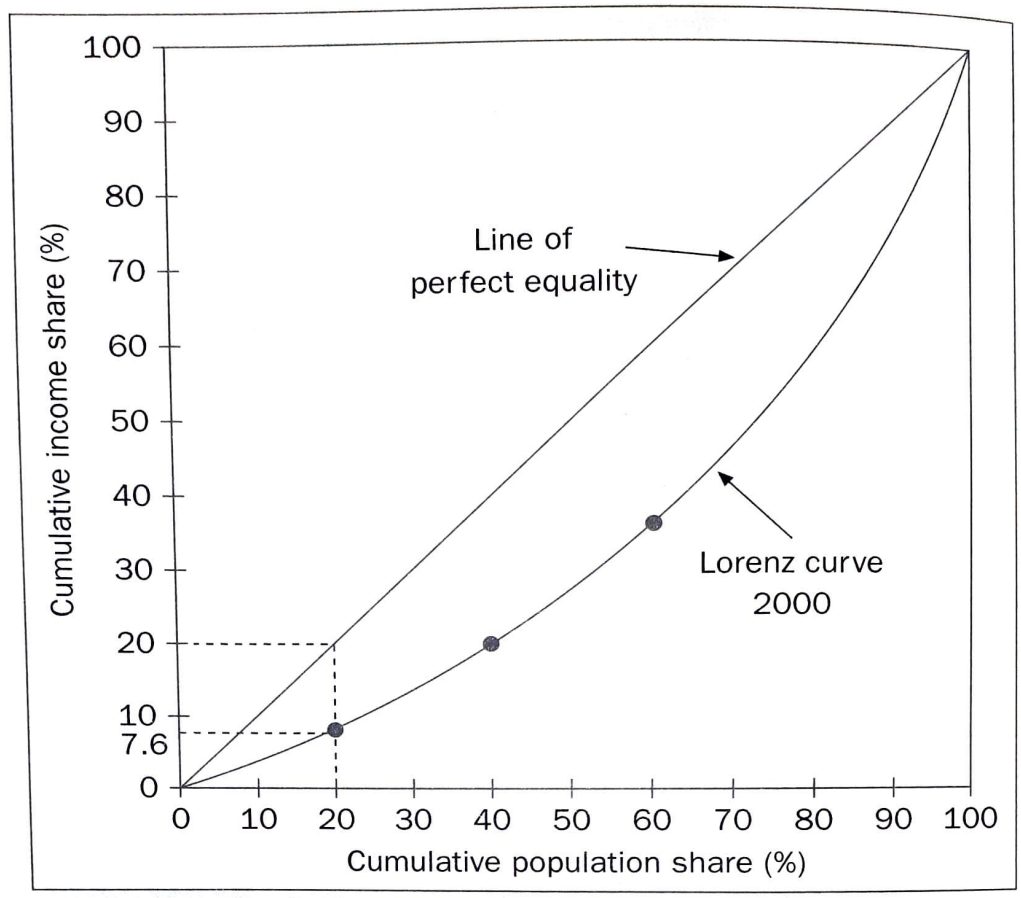
4 Comment on your results.

A response to this activity can be found on p. 459.

1.10 Lorenz curve

The **Lorenz curve** is used to contrast the *actual* distribution of a variable with the distribution which would have occurred in a situation of 'perfect equality'. The distribution of variables such as income, wealth, profit, etc. are often displayed visually using Lorenz curves.

figure 1.12
Lorenz curve



As we can see from Figure 1.12, the Lorenz curve uses *cumulative percentages* for each axis; one axis being the cumulative percentage figure for the *variable* (here income), the other being the cumulative percentage figure for the *frequency* (here people).

- The diagonal line represents *perfect equality*; for example 20 per cent of all people have 20 per cent of all incomes, and so on.
- The Lorenz curve itself represents the *actual distribution*; for example 20 per cent of all people have only 7.6 per cent of all income, and so on.
- The vertical distance between the Lorenz curve and the diagonal represents *inequality*; for example the bottom 20 per cent of income earners should have had a further 12.4 per cent of income had perfect equality existed.

The **Gini coefficient** is sometimes used with the Lorenz curve as a measure of inequality. It is the ratio of the area between the Lorenz curve and diagonal to the total area beneath the diagonal:

- If there is *perfect equality* (i.e. the plotted Lorenz curve *is* the diagonal), then the Gini coefficient is zero.
- If there is *perfect inequality* (i.e. the plotted Lorenz curve is the horizontal axis until the last person who has *all* the variable), then the Gini coefficient is one.
- The smaller the Gini coefficient, the greater the equality, and the larger the Gini coefficient the greater the inequality.

As we shall see from Self-check question 1.6 below, the above conclusions may be invalidated should different Lorenz curves intersect one another. In this case we may have to trade-off greater equality at one part of the distribution with greater inequality at another part, which may render the *overall* figure of the Gini coefficient less useful.

SELF-CHECK QUESTIONS

1.5 Draw a pie chart to display the following data. You can round your result to the nearest degree.

Road accident casualties	2000
Category	*Number killed*
Pedestrians	1,914
Pedal cyclists	323
Two-wheeled motor vehicles	942
Cars & taxis	2,019
Others	226
Total	5,424

1.6 (a) Use the following data to draw a Lorenz curve for 1979 on Figure 1.12 above (p. 20).

Cumulative per cent of income received by deciles of population.

Income receivers	*1979*	*2000*
Bottom 10%	4.7	3.0
Bottom 20%	11.0	7.6
Bottom 30%	17.5	13.9
Bottom 40%	24.2	20.3
Bottom 50%	32.2	27.5
Bottom 60%	41.2	36.2
Bottom 70%	51.8	46.7
Bottom 80%	64.3	59.2
Bottom 90%	79.6	75.0
Bottom 100%	100.0	100.0
Gini coefficient	0.248	0.337

(b) Compare the 1979 and 2000 Lorenz curves and comment on any changes in inequality.

(c) Suppose, for 2000, you were to add an extra 10.0 per cent to the incomes received by the bottom 60 per cent, 70 per cent, 80 per cent and 90 per cent, so that the respective percentage figures become 46.2, 56.7, 69.2, 85.0 (all other figures unchanged). How would this affect your interpretation of income inequality in 2000 as compared with 1979?

Note: **Answers can be found on pp. 323–324.**

REVIEW QUESTIONS

1.1 A local authority is reviewing the amenities of one of the villages in their district. It commissions a survey based on a questionnaire to investigate the age distribution (in years) of the village's population. The results of the investigation are as follows:

<5	5–15	15–25	25–35	35–45	45–55	55–65	>65
41	142	63	70	105	102	177	138

The oldest resident was 95 years old.

From the above data construct:

(a) a histogram of the data,
(b) a frequency polygon.

1.2 A survey has been conducted by the personnel department on the distances travelled to work by the production staff, with the results shown below:

Distance travelled to work (in kilometres)				
13	1	2	9	27
4	63	7	2	3
12	43	6	4	3
73	26	7	8	2
48	16	42	3	2
8	26	49	21	15
2	53	7	16	21
57	2	55	18	25
6	71	8	2	8
3	4	8	4	3

The aim of the survey is to investigate the effect on the staff of relocating the production facility. The main impact of the move will be determined by investigating the impact on those who travel: 0–5 kilometres; 5–10 kilometres; 10–30 kilometres; >30 kilometres.

Construct:

(a) a frequency table of the data,
(b) draw a histogram of the data.

1.3 A company wishes to review its distribution operation and from its time sheet records it found that 144 vehicles were loaded in a 24 hour period. A frequency distribution table was prepared from the data as follows:

Time to load (minutes)	*N° of vehicles*
40–45	4
45–50	13
50–55	17
55–60	44
60–70	59
70–80	7

From the above data construct:

(a) a cumulative frequency curve (ogive);
(b) from your ogive estimate:
  (i) the median loading time,
  (ii) the interquartile range (the difference between the upper and lower quartiles),
  (iii) the number of vehicles loaded in under 52 minutes.

1.4 The following total expenditure was incurred by a business in the last financial year:

Cost centre	*£*
Employees	1,139,410
Premises	69,004
Transport	52,524
Raw materials	83,205
Admin	23,242
Capital financing costs	71,282
Total	1,438,667

Construct a pie chart to illustrate the data.

1.5 Whilst investigating the pattern of imports into a target market, a firm collected the following data:

Year	*$ (billion)*
1991	308.962
1992	283.724
1993	305.448
1994	292.457
1995	202.743
1996	204.61
1997	226.2
1998	272.981
1999	319.01
2000	303.829

Use a vertical bar chart to illustrate the data.

1.6 The sales volume (units) by geographical area of three types of product in a particular quarter are as follows:

Model	*North*	*South*	*East*	*West*	*Export*
Galactic	4,644	4,817	4,055	4,393	1,201
Jupiter	7,884	9,735	6,468	7,089	3,067
Saturn	6,136	6,703	6,497	5,934	1,844

(a) Use a component bar chart to illustrate the data.
(b) Use a multiple bar chart to illustrate the data.

1.7 With the use of suitable graphs/charts illustrate the data in the table:

Airline	No. of passengers (millions)	
	International	Domestic
Delta	8.4	80.5
American	15.2	65.9
United	11.4	62.7
USAir	1.7	57.8
Northwest	7.6	37.9
Continental	3.0	36.9
CAAC	1.4	36.1
Al Nippon	1.7	32.8
Lufthansa	16.9	14.1
Aeroflot	3.8	25.2
British Airways	22.0	5.5
JAL	9.4	16.7
TWA	1.8	19.0
Alitalia	9.1	11.2
SAS	9.8	9.0
America West	–	15.6
Air France	13.8	1.8
Air Inter	0.1	6.8
Japan Air System	0.3	15.1
Malaysia Air	6.4	7.8

Source: ICAO

Note: Answers to Review questions can be found on pp. 376–384.

Further study and data

Texts

Bancroft, G. and O'Sullivan, G. (1993), *Quantitative methods for accounting and business studies*, 3rd edn, McGraw Hill, chapter 7.

Curwin, J. and Slater, R. (1996), *Quantitative methods for business decisions*, 4th edn, International Thompson Business Press, chapters 1 and 2.

Lawson, M., Hubbard, S. and Pugh, P. (1995), *Maths and statistics for business*, Addison Wesley Longman, chapters 1 and 2.

Morris, C. (1999), *Quantitative approaches in business studies*, 5th edn, Pitman, chapter 5.

Oakshott, L. (1996), *Essential elements of business statistics*, BPP, chapter 2.

Swift, L. (2001), *Quantitative methods for business, management and finance*, Palgrave Publishers, part DD 1–2.

Thomas, R. (1997), *Quantitative methods for business studies*, Prentice Hall, chapter 1.

Waters, D. (1997), *Quantitative methods for business*, 2nd edn, Addison Wesley Longman, chapters 2, 3 and 4.

Wisniewski, M. with Stead, R. (1996), *Foundation quantitative methods for business*, Pitman, chapters 3, 4, 5 and 6.

Sources of information and data

Of course the precise source of data you require will depend on the problem you seek to address. The following list gives a brief summary of the types of data you might find for the UK, Europe and the rest of the world in various statistical sources. Those for the UK are mainly published by the Office for National Statistics (ONS).

- **UK National Accounts** (formerly known as *National Income and Expenditure*), more normally referred to simply as the **Blue Book**. Published annually in the Autumn by the ONS, it is the most important source of data for the UK economy. Most of the data presented are annual and the tables usually cover a 20 year period.
- **UK Balance of Payments** (known as the **Pink Book**). The sister publication to the Blue Book. Published annually by the ONS.
- **Economic Trends** is published monthly by the ONS. Most of the data are quarterly, extending back over, perhaps, five years. It covers a range of areas including output, prices, employment and trade. The **Economic Trends Annual Supplement** is particularly useful for obtaining long series of both annual and quarterly data, some series going back to 1945.
- **Regional Trends** is published annually by the ONS. It includes economic and social indicators, broken down by the various regions of the UK.
- **Social Trends**, published annually by the ONS, contains data on patterns of household wealth, income and expenditure, together with data on demographic, housing and social trends.
- **Annual Abstract of Statistics**. An amalgam of topics are covered. Annual data.
- **Monthly Digest of Statistics**. As the name implies, an amalgam of statistics published monthly. It covers a wide range of topics, including economic, social and demographic.
- **General Household Survey**. The results of a sample survey of households are published annually in the GHS. Five core topics are covered – education, employment, health, housing and population and family information.
- **Labour Market Trends**. A monthly publication covering labour market issues and offering detailed information on wage rates, productivity, hours worked and so on for various sectors of the economy.
- **Business Monitor MM23**. The source of detailed information on prices.
- **Financial Statistics**. Monthly publication of the ONS relating to financial indices such as interest rates, exchange rates and the money supply.
- **Bank of England Quarterly Bulletin**. Published by the Economics Division of the Bank of England, containing a number of articles and statistics about the money supply and the financial sector.
- **National Institute Economic Review**. Unlike all of the above, this is *not* an official Government publication. However, it does contain a statistical appendix which is a very useful compilation of UK and international data on a wide range of economic issues. Published quarterly by the National Institute of Economic and Social Research.

CHAPTER 2

Central location and dispersion

Objectives

When you have read this chapter you should be able to:

- understand and use simple short-hand notation widely used in statistics;
- identify and calculate the various measures of central location, such as arithmetic mean, median and mode;
- find measures of central location for both ungrouped and grouped data;
- be aware of the circumstances under which the various measures of central location will be similar and under which they will differ;
- identify and calculate the various measures of dispersion, such as range, interquartile range, mean deviation, variance and standard deviation;
- find measures of dispersion for both ungrouped and grouped data;
- understand and interpret measures which combine aspects of central location and dispersion, such as the coefficient of variation.

Introduction

We saw in Chapter 1 that the use of tables and charts can help by giving a visual impression of patterns and trends in data. Meaning can also be given to the original (raw) data by calculating various measures of central location (average) and dispersion (spread of data around the average).

Answers to the 'Self-check questions', responses to each 'Pause for thought' and answers to the 'Review questions' can be found at the end of the book.

2.1 Notation

At least three types of average are in common use when seeking to describe raw data, namely arithmetic mean, median and mode. The formulae used in their derivation (and indeed elsewhere) make use of a type of notation with which you should become familiar.

Σ = Greek letter sigma = sum of

When applied to some variable X_i the values of the variable which are to be summed are indicated by the numbers imposed below and above the sigma sign. For example, suppose X_i refers to the scores from the throws of a dice, and we wish to sum (add together) the scores on the first three throws, we would write:

$$\sum_{i=1}^{3} X_i = X_1 + X_2 + X_3$$

$$= 3 + 5 + 1 \text{ (say)}$$

Other possibilities might include summing the scores on the second, third and fourth throws of the dice:

$$\sum_{i=2}^{4} X_i = X_2 + X_3 + X_4$$

$$= 5 + 1 + 6 \text{ (say)}$$

More generally, when we wish to sum over n values of the variable we write:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \ldots X_n$$

where n can be any number (here the number of throws of the dice).

Sometimes a particular score on the dice might occur once or more than once; in other words that particular score might occur with a *frequency* F_i. For n throws of the dice we might obtain j different scores. We would then write the total score as being given by:

$$\sum_{i=1}^{j} F_i X_i = F_1 X_1 + F_2 X_2 + F_3 X_3 + \ldots F_j X_j$$

where j is the number of different scores observed in the n throws.

For example, suppose we are faced with the following results from 12 throws of the dice:

5, 2, 6, 6, 2, 1, 6, 3, 5, 3, 2, 1

As chance would have it only five of the six possible scores on the dice have appeared (no scores of 4 obtained) in the 12 throws. Here $j = 5$ and $n = 12$.

If we let: $X_1 = 1$, occurs with frequency $F_1 = 2$
$X_2 = 2$, occurs with frequency $F_2 = 3$
$X_3 = 3$, occurs with frequency $F_3 = 2$
$X_4 = 5$, occurs with frequency $F_4 = 2$
$X_5 = 6$, occurs with frequency $F_5 = 3$

Note that if we sum the individual frequencies over the five different scores observed we get:

$$\sum_{i=1}^{5} F_i = F_1 + F_2 + F_3 + F_4 + F_5$$

$$\sum_{i=1}^{5} F_i = 2 + 3 + 2 + 2 + 3$$

$$\sum_{i=1}^{5} F_i = 12$$

So the sum of the individual frequencies for each of the five different scores equals the total number of observations, n, which in this case equals 12.

As already mentioned we only have five different numbers, so $j = 5$ in our earlier formula:

$$\sum_{i=1}^{5} F_i X_i = F_1X_1 + F_2X_2 + F_3X_3 + F_4X_4 + F_5X_5$$
$$= (2 \times 1) + (3 \times 2) + (2 \times 3) + (2 \times 5) + (3 \times 6)$$
$$= 2 + 6 + 6 + 10 + 18$$

$$\sum_{i=1}^{5} F_i X_i = 42$$

You will come across the use of this shorthand notation involving the sigma sign throughout this chapter and in subsequent chapters. It will therefore be helpful to practice *using* the notation before going any further.

SELF-CHECK QUESTIONS

2.1 Expand (write out fully) each of the following expressions:

(a) $\sum_{i=1}^{4} X_i$ **(b)** $\sum_{i=2}^{5} X_i$ **(c)** $\sum_{i=1}^{3} F_i X_i$ **(d)** $\sum_{i=3}^{6} F_i X_i$

2.2 Write out each of the following expressions using sigma notation:

(a) $X_1 + X_2 + X_3 + X_4 + X_5$
(b) $X_5 + X_6 + X_7$
(c) $F_1X_1 + F_2X_2 + F_3X_3 + F_4X_4$
(d) $F_2X_2 + F_3X_3 + F_4X_4$

2.3 The following scores are obtained for 15 throws of a dice:

5, 6, 4, 3, 6, 5, 1, 1, 2, 4, 3, 1, 2, 4, 6

Find $\sum_{i=1}^{j} F_i X_i$

where j **= the number of different scores observed**
F_i **= frequency of a particular score**
X_i **= particular score**

Note: **Answers can be found on p. 325.**

2.2 Measures of central location

A number of measures of central location or 'average' are widely used in an attempt to give meaning to raw data. Here we consider the arithmetic mean, median and mode for both *ungrouped data* (all individual items known) and *grouped data* (items only identified within class intervals).

The arithmetic mean

This is the simple average of everyday use, and is often represented by the symbol $\bar{X}$.

Ungrouped data

Where individual or *ungrouped* data are available, the following formula is commonly used:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where $\bar{X}$ = arithmetic mean
X_i = value of each item of data
n = number of items of data

WORKED EXAMPLE **2.1**

Suppose the daily takings of a small corner shop are as shown below:

	Mon.	*Tue.*	*Wed.*	*Thur.*	*Fri.*
Daily sales	£620	£660	£600	£480	£710

Solution

Using the formula:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{5} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5} = \frac{620 + 660 + 600 + 480 + 710}{5}$$

$$= \frac{3{,}070}{5} = £614$$

The average daily sales are £614 and this represents a typical figure around which the rest of the data will cluster.

Grouped data

More usually, as we noted in Chapter 1, data are *grouped* into a frequency table with various class intervals. To deal with such data the simplifying assumption must be made that *within any given class interval* the items of data fall on the *class mid-point*. This is equivalent to assuming that the items of data are evenly spread within any given class interval.

We may now make use of the following formulae:

$$\bar{X} = \frac{\sum_{i=1}^{j} F_i X_i}{\sum_{i=1}^{j} F_i} = \frac{F_1 X_1 + F_2 X_2 + \ldots F_j X_j}{F_1 + F_2 + \ldots F_j}$$

where F_i = frequency of ith class interval
X_i = mid-point of ith class interval
j = number of class intervals

WORKED EXAMPLE **2.2**

Suppose that a survey of the prices of 60 items sold in a shop gives the results below (five class intervals for prices):

Price of item (£)	*Number of items sold*
1.5–2.5	15
2.5–3.5	2
3.5–4.5	19
4.5–5.5	10
5.5–6.5	14

Solution

We can use our formula for grouped data to calculate the average (arithmetic mean) price of the items sold (see Table 2.1).

table 2.1
The average (arithmetic mean) price of items sold

Price of items (£)	*Class mid-points* X_i	*Number of items sold* F_i	F_iX_i
1.5–2.5	2	15	30
2.5–3.5	3	2	6
3.5–4.5	4	19	76
4.5–5.5	5	10	50
5.5–6.5	6	14	84
		$\sum_{i=1}^{5} F_i = 60$	$\sum_{i=1}^{5} F_iX_i = 246$

$$\bar{X} = \frac{\sum_{i=1}^{5} F_iX_i}{\sum_{i=1}^{5} F_i} = \frac{246}{60} = 4.1$$

In other words the average price of the items sold in the shop was £4.10 pence.

DID YOU KNOW?

The mean time for a person speaking each day is less than 12 minutes, and for a doubling of paper documents held on file by businesses is less than 31 years (despite IT!)

The median

This is that value which divides the data set into two equal halves; 50 per cent of values lying below the median, and 50 per cent of values lying above the median.

The approach we consider below uses *arithmetical* methods for estimating the median value for both grouped and ungrouped data.

Ungrouped data

To find the median:

(i) Construct an *array* (i.e. place the data in numerical order – whether rising or falling).

(ii) Find the median *position* $\frac{n+1}{2}$ where n = number of values.

(iii) Find the median *value*.

WORKED EXAMPLE **2.3**

Find the median value of machine output over a five day period:

Daily output				
310	340	360	320	330

Solution

Step 1 Place in an array 310 320 330 340 360.

Step 2 Find the median position using the equation $\frac{n+1}{2}$ where n = the number of values. Median position = $\frac{5+1}{2}$ = third item.

Step 3 Read the value of the third item in the array, i.e. 330 units.

Note: Where there is an *even* number of items there will be *two* middle items. If so, take the *average* of these two middle items.

Grouped data

The following steps can be used to calculate arithmetically the median value for data presented in a frequency table:

- Find the median *position*.
- Find the *class-interval* in which the median observation lies.
- Assume that all items in this class-interval are equally spaced.
- Estimate the median.

WORKED EXAMPLE **2.4**

Find the median height of students from the data given in Table 2.2 (p. 34).

Solution

$$\text{Median position} = \frac{n+1}{2}$$

$$= \frac{20+1}{2}$$

$$= 10.5$$

The class interval in which this median position lies is 170–175.

table 2.2
The height of students

X_i Heights (cm)	F_i Frequency (number of students)
150 and under 155	1
155 and under 160	1
160 and under 165	2
165 and under 170	3
170 and under 175	6
175 and under 180	2
180 and under 185	4
185 and under 190	1
	20

The median value can be found using the formula:

$$\text{LCB} + \text{class width} \times \frac{\text{number of observations to median position}}{\text{total number of observations in median class interval}}$$

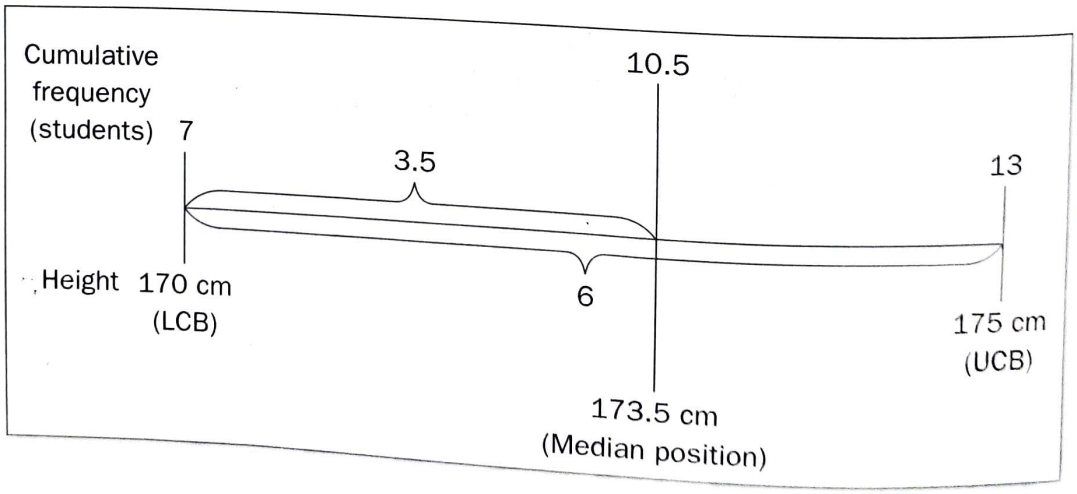
Where LCB = lower class boundary (of median class interval).

$$\text{Median} = 170 \text{ cm} + \left(5 \text{ cm} \times \frac{3.5}{6}\right)$$

$$\text{Median} = 172.9 \text{ cm}$$

Figure 2.1 gives a visual picture of the use of our formula in finding the median height.

figure 2.1
Median class interval for table 2.2



The class interval in which the median position falls has a lower class boundary (LCB) of 170 cm and an upper class boundary (UCB) of 175 cm, giving a *class width* of 5 cm. Now the cumulative frequency up to the LCB is seven students, which is less than the median position which occurs with a cumulative frequency of 10.5 students.

Clearly the median value is larger than 170 cm as we wish to go a further 3.5 students (observations) to reach the median position. However there are six students assumed to be equally spaced along this class interval of 5 cm. We therefore wish to go 3.5/6ths of the way along the 5 cm interval in order to find the median height. This gives the formula:

$$170 \text{ cm} + (5 \text{ cm} \times \frac{3.5}{6}) = 172.9 \text{ cm}$$ which we noted above.

We have already noted in Chapter 1 (p. 12) how the use of *cumulative frequency curves* (or ogives) can help find the median value by *graphical* methods.

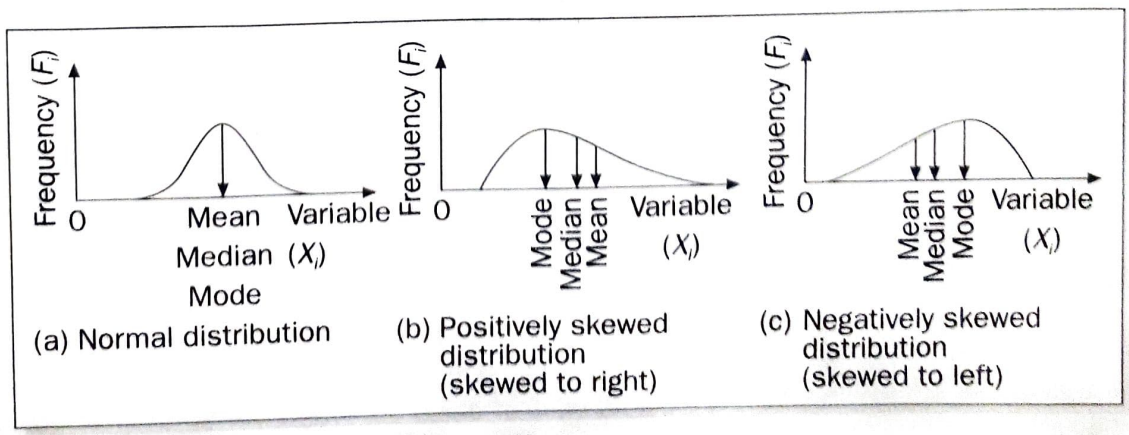
The mode

This is that value which occurs with greatest frequency. When data are grouped, then the *class interval* with the highest frequency is referred to as the modal class interval. In Table 2.2 above the class interval 170–175 cm is the modal class interval, with six students in the interval.

2.3 Normal and skewed distribution

When the set of data is distributed in a perfectly symmetrical way, as in Figure 2.2(a), then all three types of average have the same value. Such a symmetrical distribution is often referred to as a **normal distribution**. However when, as is more usually the case, the set of data is skewed in one direction or another, as in Figures 2.2(b) and (c), then the three types of average will cease to be identical. In fact the arithmetic mean will always be most heavily influenced by the direction of skew, i.e. the direction being described by the tail of the distribution, which represents the side of the distribution with fewest observations. In other words, the arithmetic mean will be most affected by a *few* extreme values, whether higher values (skewed to the right) or lower values (skewed to the left).

figure 2.2
Normal and skewed frequency distributions and measures of central location



When data are skewed to the right or left, you must be careful in interpreting any loose talk of 'average' values. For example if the distribution of wages in a firm is skewed to the right, as in Figure 2.2(b), the employers might claim an 'average' wage of, say, £180 per week and the trade union an 'average' wage of, say, £160 a week and both might be right! However the employers may be selecting the type of 'average' (arithmetic mean) most favourable to showing them as a high paying company and the union may be selecting the type of average (median) most favourable as a basis for pressing for higher wages.

PAUSE FOR THOUGHT 2.1 *In what direction would you expect the UK income distribution to be skewed? How might this affect median and mean incomes?*

WORKED EXAMPLE **2.5**

The following distribution shows the weekly output of production workers in the company:

table 2.3
Weekly output of production workers

Output (units)	*Number of employees*
100–160	1
160–180	5
180–200	10
200–220	35
220–240	55
240–260	74
260–300	20
	200

(i) Find the arithmetic mean of the weekly output.
(ii) Find the median weekly output.
(iii) Why do (i) and (ii) differ?

Solution

(i) Since we are dealing with grouped data we must take *class mid-points* for the variable X_i. We then use the formula:

$$\bar{X} = \frac{\sum_{i=1}^{j} F_i X_i}{\sum_{i=1}^{j} F_i}$$

where F_i = class frequency
X_i = class mid-point
j = number of class intervals

X_i (class mid-points)	F_i (class frequency)	F_iX_i
130	1	130
170	5	850
190	10	1,900
210	35	7,350
230	55	12,650
250	74	18,500
280	20	5,600
	$\sum_{i=1}^{7} F_i = 200$	$\sum_{i=1}^{7} F_iX_i = 46{,}980$

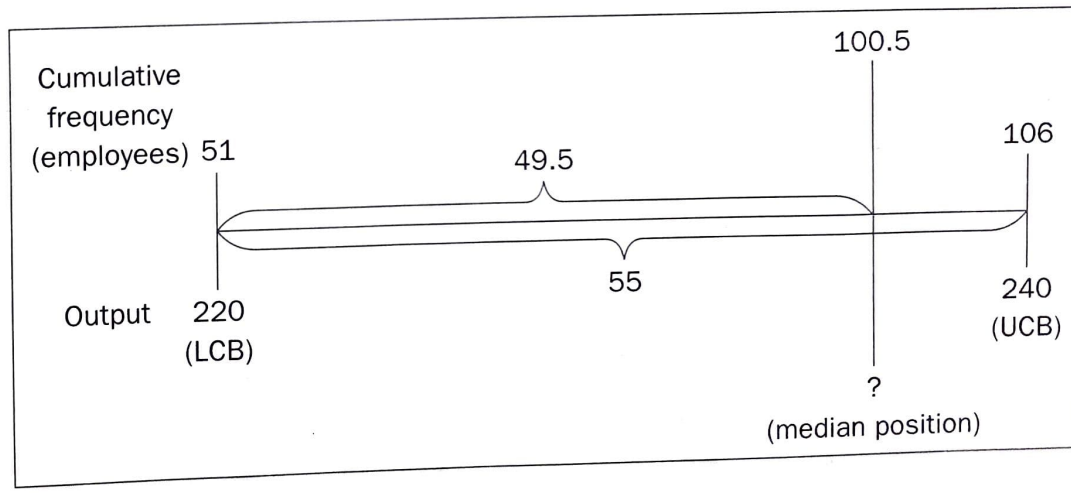
$$\bar{X} = \frac{\sum_{i=1}^{7} F_iX_i}{\sum_{i=1}^{7} F_i} = \frac{46{,}980}{200} = 234.9$$

The arithmetic mean of the weekly output is 234.9 units.

(ii) The median position $= \frac{n+1}{2} = \frac{201}{2} = 100.5$ employees.

The median class interval for output is 220–240 units, as shown in Figure 2.3.

figure 2.3
Finding the Median Value



Using our formula for the median, and assuming an equal spacing of the 55 employees along the median class interval of 20 units of output, we have:

$$\text{LCB} + \text{class width} \times \frac{\text{number of observations to median position}}{\text{total number of observations in class interval}}$$

i.e.

$$\text{Median} = 220 + \left(20 \times \frac{49.5}{55}\right) = 238 \text{ units}$$

(iii) The arithmetical mean output (234.9 units) is lower than the median output (238 units). We would expect this to be the case since the data are clearly skewed to the left. It is similar to Figure 2.2(c) and the arithmetic mean (simple average) will be pulled down by the few extremely low values.

PAUSE FOR THOUGHT 2.2 *Before calculating the solution to Self-check question 2.6 below, look briefly at the table of data and decide on the direction of skew and the likely impact this will have on the mean and median values.*

SELF-CHECK QUESTIONS

2.4 Passenger cars produced for export (X_i) in thousands in each month during the past year were as follows:

69.3	75.5	76.3	94.4	88.7	77.2	86.4	75.5	79.0	83.3	82.2	78.1

(a) What is the value of X_i when:
- **(i) $i = 4$**
- **(ii) $i = 7$**
- **(iii) $i = 10$**

(b) What is the mean monthly production of cars?

(c) What is the mean monthly production of cars over the final six months?

(d) What is the mean monthly production of cars during months 4–9 inclusive?

2.5 The following figures show the annual salaries in £s of 20 workers in a small firm. Calculate the arithmetic mean, median and mode salary and comment on your results:

15,180	19,870	14,375	15,767	15,870	15,180	14,375	36,938
15,180	46,132	15,525	19,600	14,375	23,069	16,767	16,767
17,880	14,375	14,375	14,375				

2.6 An investment analyst receives the following table of data showing the percentage changes in labour costs of senior managers in a multi-national company over a 12 month period. The highest change observed was + 172 per cent for a senior manager in one of the plants in Latin America.

Percentage change	*Frequency*
−5 to under 0	2
0 to under 5	32
5 to under 10	25
10 to under 15	10
15 to under 20	8
20 to under 25	3
25 to under 30	2
30 to under 35	5
35 to under 40	4
40 to under 100	3
100 to 172	4

(a) What is the mean annual percentage change in labour costs for senior managers in this company?
(b) What is the estimate of the median value?
(c) What is the modal class interval?

Note: **Answers can be found on pp. 325–328.**

2.4 Measures of dispersion

As well as measures of central tendency or average, it is helpful to have a measure of the extent of **dispersion** or spread around that average. The formulae and brief outlines for the most widely used measures of dispersion are presented here.

As we can see from Figure 2.4, two distributions showing the prices of items sold in a shop may have similar measures of central location but may be very different in terms of dispersion around any 'average'. Clearly distribution B is more widely dispersed than distribution A around the same arithmetic mean $\bar{X}$.

- **Range** Perhaps the simplest measure of dispersion is to take the absolute difference between the highest and lowest value of the raw data. In Figure 2.4 the range for distribution A is £30, but the range for distribution B is higher at £50.
- **Interquartile range** This is the absolute difference between the upper and lower quartiles of the distribution. We saw how to calculate these quartiles in Chapter 1, p. 12.

Interquartile range = upper quartile – lower quartile

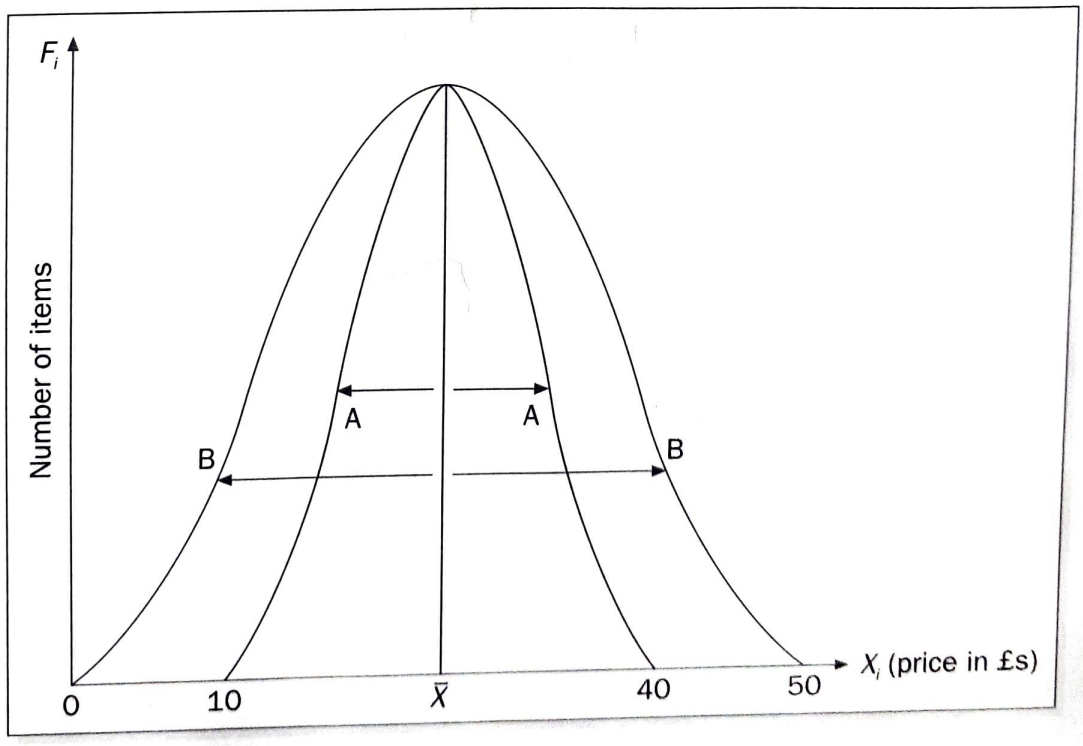


figure 2.4
Same mean ($\bar{X}$), different dispersion for two distributions A and B

▶ **Semi-interquartile range** This is half the interquartile range. It is sometimes called the *quartile deviation.*

$$\text{Semi-interquartile range} = \frac{\text{upper quartile} - \text{lower quartile}}{2}$$

All the measures of dispersion considered so far have involved comparing two different points on the respective frequency distributions, such as the maximum and minimum points (the range) or the upper and lower quartiles (the interquartile range). We now turn to more useful measures of dispersion which compare *all* the points on the respective frequency distributions.

DID YOU KNOW?

The mean income for the top fifth of UK households is more than 20 times higher than that for the bottom fifth of UK households, before *the impact of the tax and benefit system.* After *the impact of taxes and benefits, the respective mean incomes only differ by a factor of less than three.*

Mean deviation

This is the average of the absolute deviations from the arithmetic mean, ignoring the sign. Of course if we did *not* ignore the sign, the average deviation from the arithmetic mean would always be zero! When two straight lines (rather than curved brackets) surround a number or variable it is referred to as the *modulus* and means ignore the sign.

The mean deviation is only rarely used as a measure of dispersion. However it is worth some consideration as it is the basis for the more widely used measures of variance and standard deviation.

Ungrouped data

Suppose three throws of a dice yield the following values:

$X_1 = 2,\ X_2 = 4,\ X_3 = 3$

Clearly the arithmetic mean, $\bar{X}$, is 3. To find the **mean deviation** (MD) we need to find the average deviation from the mean, ignoring the sign:

$$\text{MD} = \frac{|X_1 - \bar{X}| + |X_2 - \bar{X}| + |X_3 - \bar{X}|}{n}$$

$$\text{MD} = \frac{|2 - 3| + |4 - 3| + |3 - 3|}{3} = \frac{|-1| + |+1| + |0|}{3} = \frac{2}{3}$$

On average, each score on the three throws of the dice is $\frac{2}{3}$ of a unit away from the average score of 3.

Mean deviation with ungrouped data

$$\text{More generally MD} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

where $|\ |$ = modulus (i.e. ignore the sign)
n = number of observations

Grouped data

If we return to our example of the prices of 60 items bought in a shop (Table 2.1), we can now use the mean deviation to give a measure of dispersion. We noted (p. 32) that the average price ($\bar{X}$) was £4.1 per item sold. We now apply the formula for mean deviation with grouped data:

Mean deviation with grouped data

$$\text{MD} = \frac{\sum_{i=1}^{j} F_i|X_i - \bar{X}|}{\sum_{i=1}^{j} F_i}$$

where F_i = class frequency
X_i = class mid-point
$\bar{X}$ = arithmetic mean
j = number of class intervals
$|\ |$ = modulus (i.e. ignore sign)

PAUSE FOR THOUGHT 2.3 *How might you expect the mean deviation for the age of the UK population to change over time?*

WORKED EXAMPLE **2.6**

Find the mean deviation for the data shown in Table 2.1 (p. 32).

Solution

Price of item (£)	*Class mid-point* X_i	*Number of items* F_i	$\lvert X_i - \bar{X}\rvert$	$F_i\lvert X_i - \bar{X}\rvert$
1.5–2.5	2	15	|−2.1|	|−31.5|
2.5–3.5	3	2	|−1.1|	| −2.2|
3.5–4.5	4	19	|−0.1|	| −1.9|
4.5–5.5	5	10	| 0.9|	| 9.0|
5.5–6.5	6	14	| 1.9|	| 26.6|
		60		71.2

Note: $\bar{X} = 4.1$.

$$\text{MD} = \frac{\sum_{i=1}^{5} F_i|X_i - \bar{X}|}{\sum_{i=1}^{5} F_i} = \frac{71.2}{60} = 1.19$$

The average deviation from the arithmetic mean price is £1.19 per item.

Variance

If we square all the deviations from the arithmetic mean, then we no longer need to bother with the signs since all the values will be positive. We can then replace the straight line brackets (modulus) for the mean deviation with the more usual round brackets.

The resulting measure is now called the **variance**, which can then be regarded as the average of the *squared* deviations from the mean. Note that all the units are squared, e.g. square pounds, square metres, etc.

Ungrouped data

Variance for ungrouped data

$$\text{Variance } (s^2) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

WORKED EXAMPLE **2.7**

Suppose we have to find the variance of the following eight items of raw data.

8, 10, 12, 14, 16, 18, 20, 22

Solution

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{120}{8} = 15$$

X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
8	(−7)	49
10	(−5)	25
12	(−3)	9
14	(−1)	1
16	(1)	1
18	(3)	9
20	(5)	25
22	(7)	49
		168

$$s^2 = \frac{\sum_{i=1}^{8}(X_i - \bar{X})^2}{n} = \frac{168}{8} = 21 \text{ square units}$$

Variance (s^2) = 21 square units

Grouped data

The formula for calculating the variance when data are grouped is as follows:

Variance for grouped data

$$\text{Variance } (s^2) = \frac{\sum_{i=1}^{j} F_i(X_i - \bar{X})^2}{\sum_{i=1}^{j} F_i}$$

where F_i = class frequency
X_i = class mid-point
$\bar{X}$ = arithmetic mean
j = number of class intervals

However it can be shown (see p. 47) that this formula simplifies to the following:

Variance for grouped data: operational formula

$$\text{Variance } (s^2) = \frac{\sum_{i=1}^{j} F_i X_i^2}{\sum_{i=1}^{j} F_i} - \left(\frac{\sum_{i=1}^{j} F_i X_i}{\sum_{i=1}^{j} F_i} \right)^2$$

$$\text{Variance } (s^2) = \frac{\sum_{i=1}^{j} F_i X_i^2}{\sum_{i=1}^{j} F_i} - (\bar{X})^2$$

This is a useful formula for finding the *variance* since we will already have found the second term when calculating the *arithmetic mean*.

WORKED EXAMPLE **2.8**

Let us apply this formula to our earlier problem involving prices of items sold in Worked Example 2.6.

Price of item (£)	*Class mid-points* X_i	*Number of items* F_i	F_iX_i	$F_iX_i^2$
1.5–2.5	2	15	30	60
2.5–3.5	3	2	6	18
3.5–4.5	4	19	76	304
4.5–5.5	5	10	50	250
5.5–6.5	6	14	84	504
		60	246	1,136

$$s^2 = \frac{\sum F_iX_i^2}{\sum F_i} - \left(\frac{\sum F_iX_i}{\sum F_i} \right)^2$$

$$s^2 = \frac{1{,}136}{60} - \left(\frac{246}{60}\right)^2 = 18.9 - (4.1)^2$$

$s^2 = 2.1$ square pounds

Note: $\sum_{i=1}^{j} F_i X_i^2 = F_1 X_1^2 + F_2 X_2^2 + \ldots F_j X_j^2$

In other words the square is only on the X_i term and *not* on $F_i X_i$.

We can obtain the column $F_i X_i^2$ by *either* multiplying the $F_i X_i$ column by the corresponding X_i *or* by multiplying the F_i column by the corresponding X_i^2. Check this yourself in the table above.

Standard deviation

Rather than use square units, it is more realistic to express solutions in terms of single units. The **standard deviation** does this by taking the square root of the variance.

The standard deviation (s) is then the square root of the average of the squared deviations from the mean. The following formula applies to both ungrouped and grouped data:

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

$$\text{i.e.} \qquad s \qquad = \sqrt{s^2}$$

WORKED EXAMPLE **2.9**

We can use our earlier data on 20 student heights to illustrate the calculation of both the variance and the standard deviation as measures of dispersion.

Solution

Heights (cm)	*Number* F_i	*Mid-points* X_i	$F_i X_i$	$F_i X_i^2$
150 and under 155	1	152.5	152.5	23,256.25
155 and under 160	1	157.5	157.5	24,806.25
160 and under 165	2	162.5	325	52,812.50
165 and under 170	3	167.5	502.5	84,168.75
170 and under 175	6	172.5	1,035	178,537.50
175 and under 180	2	177.5	355	63,012.50
180 and under 185	4	182.5	730	133,225.00
185 and under 190	1	187.5	187.5	35,156.25
	20		3,445	594,975

$$\bar{X} = \frac{\sum_{i=1}^{8} F_i X_i}{\sum_{i=1}^{8} F_i} = \frac{3{,}445}{20} = 172.25 \text{ cm}$$

$$s^2 = \frac{\sum_{i=1}^{8} F_i X_i^2}{\sum_{i=1}^{8} F_i} - \left(\frac{\sum_{i=1}^{8} F_i X_i}{\sum_{i=1}^{8} F_i} \right)^2$$

$$s^2 = \frac{594{,}975}{20} - \left(\frac{3{,}445}{20} \right)^2$$

$s^2 = 29{,}748.75 - 29{,}670.06$

s^2 = Variance = 78.69 square cm

s = Standard deviation = $\sqrt{78.69}$

i.e. $s = 8.87$ cm

ACTIVITY **2**

You have been commissioned by a manufacturing company to examine the potential of the Greek market. Part of the investigation requires work on the demography of Greece. Use a spreadsheet with which you are familiar to derive the following information and data. Here the formulae and functions are given in terms of *Microsoft Excel*.

1 *Preparation*

Type the following data and headings exactly as shown below

	A	B	C	D
1	**Greece: Population 1986–1999**			
2				
3	Area (sq km)	131,908		
4				
5		**population**	**% change**	**Pop. density**
6				
7	1986	9,642,505		
8	1987	9,729,350		
9	1988	9,789,513		
10	1989	9,846,627		
11	1990	9,895,801		
12	1991	9,934,249		
13	1992	9,963,604		
14	1993	9,983,490		
15	1994	10,004,401		
16	1995	10,038,672		
17	1996	10,088,700		
18	1997	10,200,000		
19	1998	10,311,300		
20	1999	10,422,600		

2 *Calculate the percentage change in population*

(a) Create a formula in **C8** (*not* **C7**) to give the percentage change of population from 1986 to 1987.

(b) Copy this formula down the column in order to calculate the percentage change for all the years up to and including 1999.

The formula in **C8** needs to be a relative reference (cell references that adjust when copied to new locations).

3 Calculate the population density

A In cell **D7** create a formula, by referencing cell **B3** to calculate the population per sq. km (population density). Before this formula is copied down make cell **B3** an absolute reference. Cell references in formulae must be made absolute if they are not to be adjusted when copied. To create an absolute cell reference in *Excel* do the following:

1 enter the formula in the usual way;

2 position the cursor within the formula bar to the right of the cell reference that is to be made absolute (it will change from a cross to an 'I' shape);

3 press the F4 key once and a $ sign will appear with the cell reference ($B$3).

B Copy the formula down the column.

A formula containing an absolute cell reference can be copied anywhere in the spreadsheet and it will continue to refer back to the absolute cell (in this case **B3**).

4 Construct a graph

Construct a graph showing both the population and the percentage change in population.

5 Calculate the mean percentage change in population for the period.

6 Estimate the future population

(a) Use the mean percentage change in population to estimate the population for 2000, 2001 and 2002.

(b) Put these projections on to your graph.

A response to this activity can be found on p. 461.

2.5 Coefficient of variation (CofV)

This is a widely used measure of *relative dispersion*. It relates an *absolute* measure of dispersion (the standard deviation) to the *absolute* value of the arithmetic mean around which the dispersion takes place. Clearly a data set A with a standard deviation (s) of ten units and mean ($\bar{X}$) of ten units has a *greater* relative dispersion than a data set B with a higher absolute standard deviation of 20 units but a still higher absolute mean ($\bar{X}$) of 50 units.

We would say that:

$$\text{C of } V_A = \frac{s}{\bar{X}} = \frac{10}{10} = 1.0$$

$$\text{C of } V_B = \frac{s}{\bar{X}} = \frac{20}{50} = 0.4$$

The coefficient of variation can be expressed (as here) as a decimal, or as a percentage. To obtain a percentage we simply multiply by 100, giving 100 per cent and 40 per cent for this example. The data set with the highest coefficient of variation has the greatest relative dispersion.

Proof of operational formula for variance

The formula for the variance (p. 43) can be made more useful by simplifying, as follows:

$$s^2 = \frac{\sum F_i(X_i - \bar{X})^2}{\sum F_i}$$

$$s^2 = \frac{\sum F_i(X_i^2 - 2X_i\bar{X} + \bar{X}^2)}{\sum F_i} \text{ (expanding the bracket)}$$

$$s^2 = \frac{\sum F_i X_i^2}{\sum F_i} - 2\bar{X}\frac{\sum F_i X_i}{\sum F_i} + \frac{\bar{X}^2 \sum F_i}{\sum F_i}$$

$$s^2 = \frac{\sum F_i X_i^2}{\sum F_i} - 2\bar{X}^2 + \bar{X}^2$$

$$s^2 = \frac{\sum F_i X_i^2}{\sum F_i} - \bar{X}^2$$

$$s^2 = \frac{\sum F_i X_i^2}{\sum F_i} - \left(\frac{\sum F_i X_i}{\sum F_i}\right)^2$$

SELF-CHECK QUESTIONS

2.7 Look back to the data in Self-check question 2.4 (p. 38). Find the variance and standard deviation for monthly production of passenger cars for export.

2.8 Look back to the data in Self-check question 2.6 (p. 38). Find the variance and standard deviation for the percentage change in labour costs of senior managers throughout the multinational company.

Note: **Answers can be found on p. 328.**

REVIEW QUESTIONS

2.1 A frozen food company has a contract with Farm A, to provide peas for their processing plant. The daily supply of peas (in tonnes) over the past two weeks from Farm A has been as follows:

Farm A

357.38	262.80	319.95	412.90	398.46	330.33	341.27
329.33	332.04	309.42	229.88	259.43	337.99	383.31

The frozen food company also contracts with Farm B, The daily supply of peas (in tonnes) over the past two weeks from Farm B has been as follows:

Farm B

364.32	295.18	352.92	380.14	314.39	343.87	290.43
291.21	306.03	348.88	333.81	291.99	324.79	315.25

Calculate the respective values for each farm in terms of:

(a) the mean;
(b) the range;
(c) the interquartile range;
(d) the semi-interquartile range.

Use your calculations to compare and contrast the relative contributions of each farm to the frozen food company.

2.2 Use the data for Farm A and Farm B in question 2.1 above to calculate the mean deviations for both data sets. What do your results show?

2.3 A company requires that chilled food cabinets in its supermarkets must maintain an average hourly temperature of 3.75°C ± 0.5°C. The manager at one of the supermarkets suspects that the performance of one of the shop's cabinets fails to meet this standard and therefore decides to monitor its performance hourly over a 30 day period with the following results:

Temperature (°C)	***Frequency***
0–1	1
1–2	11
2–3	123
3–4	322
4–5	223
5–6	39
6–7	1

Find the arithmetic mean hourly temperature and the mean deviation (for grouped data) to assess whether the equipment conforms to the company's policy.

2.4 The number of visitors, in thousands, to an amusement park over a summer season were recorded as follows:

Visitors (000)	***Frequency***
6–8	4
8–10	17
10–12	61
12–14	47
14–16	18
16–18	3

Calculate the mean attendance during the summer season and the variance and standard deviation for that attendance. Comment on your results.

2.5 A new production line had a mean daily rejection rate of 196 units with a standard deviation of 21.48 units in the first three months of operation. In the next three months the mean daily rejection rate was 94 units with a standard deviation of 12.62 units. Find the coefficient of variation in each case and use it to comment on the relative dispersion of the production line over the two time periods.

2.6 A brewing company wishes to launch a new canned lager. It has close links with a major supermarket chain which will only permit a promotion of the lager in two of its stores. The selected stores monitor their sales of all brands of canned lager over a weekend period with the following results, which were communicated to the brewery:

Sales value of lager (£)	*Frequency*	
	Store A	*Store B*
0–2.5	27	1
2.5–5	114	3
5–7.5	333	31
7.5–10	530	142
10–12.5	504	328
12.5–15	334	498
15–17.5	121	504
17.5–20	29	351
20–22.5	5	110
22.5–25	2	29
25–27.5	1	3
	2,000	2,000

What advice would you give the management of the brewery on the differences and similarities in the pattern of expenditure in stores A and B. You may wish to consider:

(a) Illustrating the data graphically – using the same axes for each store.
(b) The mean consumer expenditure in each store.
(c) The variability of customer expenditure.
(d) Which supermarket provides the best opportunity for a successful promotion of the new lager.

Answers to Review questions can be found on pp. 384–392.

Further study and data

Texts

Bancroft, G. and O'Sullivan, G. (1993), *Quantitative methods for accounting and business studies*, 3rd edn, McGraw Hill, chapter 8.

Curwin, J. and Slater, R. (1996), *Quantitative methods for business decisions*, 4th edn, International Thompson Business Press, chapters 3 and 4.

CHAPTER 3

Regression and correlation

Objectives

By the end of this chapter you should be able to:

- calculate the unique 'line of best fit' (**least squares line**) between a dependent variable (Y) and an independent variable (X). In other words you should be able to find the equation of that unique line $Y = mX + c$ which minimises the sum of squared deviations from the line. This is sometimes called the linear regression line.
- calculate the **coefficient of determination (R^2)** and Pearson's **coefficient of correlation (R)** between the two variables, thereby helping to establish the 'goodness of fit' of your least squares line.
- calculate and interpret Spearman's '**coefficient of rank correlation**'. This seeks to establish whether or not there is a strong linear relationship between two sets of ranked data.
- appreciate the approach to finding the regression line (linear or non-linear) which best fits more than two variables. In other words be broadly familiar with the techniques and problems of **multiple regression analysis**.

Introduction

Business activity is an aggregate of tens of thousands of decisions taken daily by small and large firms alike. Decisions bring about change and, hopefully, such change will benefit the firm and the wider economy. For example, 'If I spend more on advertising, sales volume will grow', 'If I invest in the latest equipment, productivity will increase'.

The cornerstone of many of these expectations is usually past experience by the firm in question, or the widely reported experiences of other firms. Of course we can be more confident that the expected outcome will in fact occur, if it is supported by the recorded experiences of a *large number* of decision takers.

The statistical techniques of regression and correlation are important in helping determine whether such relationships can be established. Usually the firm is interested in whether changing some independent variable (say advertising) will have the desired effect on some dependent variable (say sales volume). As we shall see, the section on **regression** will help us identify the equation of the line or curve which 'best fits' the recorded data, often displayed in the form of a scatter diagram. The section on **correlation** will help us determine 'how well' that line or curve actually fits that recorded data.

Throughout this chapter we shall present two alternative approaches to calculation. The first will use the *original data* for the dependent and independent variables, Y and X respectively. The second will use *redefined data* for these variables, giving us the opportunity to use much simpler 'coding formulae' for our calculations. Although both approaches will be presented, you can concentrate on that approach followed by your lecture course (or the approach you find the easiest!).

Answers to the 'Self-check question', responses to each 'Pause for thought' and answers to the 'Review questions' can be found at the end of the book.

3.1 Regression analysis

Regression analysis involves establishing a relationship between two or more variables. Here we initially illustrate in terms of a *linear* relationship, although more sophisticated approaches can establish similar *non-linear* relationships. In 'simple' regression analysis we assume two variables only, Y the dependent variable and X the independent variable.

Simple linear regression

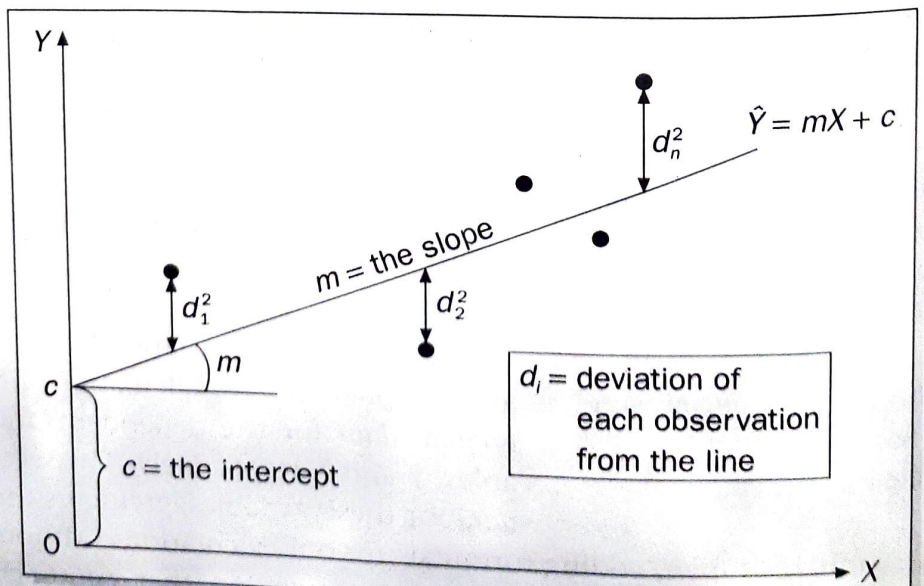
We consider the equation of a straight line (linear equation) in more detail in Chapter 12 (p. 306). Here we note that the general equation of a straight line is:

$$Y = mX + c$$

where Y = dependent variable
X = independent variable
m = gradient of the line
c = point where the line intersects the Y axis

Figure 3.1 shows a scatter diagram with dots representing the different co-ordinates (X,Y) plotted on the diagram. Our aim is to find an estimated line ($\hat{Y} = mX + c$) which best fits the data. We use the symbol $\hat{Y}$ to refer to the *estimated* value of Y from the least squares line.

figure 3.1
Finding the least squares line (LSL) $\hat{Y} = mX + c$



One possibility would be to find that line which minimises the sum of deviations (d_i) of each observation from the line. However some observations would have positive deviations (above the line) and some would have negative deviations (below the line). We would then be faced with the same problem noted in Chapter 2 for the mean deviation (p. 40) of identifying these signs ± and then ignoring them (i.e. taking the *modulus*). It is much easier to *square* all these deviations so that all the signs will be positive.

We shall therefore use the least squares line as our line of best fit, i.e. that (straight) line which minimises the sum of the squared deviations of the actual observations from the line.

In other words the least squares line is that (unique) line for which:

$d_1^2 + d_2^2 + \ldots d_n^2$ is a minimum.

In terms of Figure 3.1, the least squares line will be such that $\sum_{i=1}^{n} d_i^2$ is a minimum.

The slope(m) and intercept (c) of this unique line can be found using *either* of the following two approaches. In each case the formula outlined is obtained via differentiating expressions in order to find those values of m and c which will minimise the sum of squared deviations from the (straight) line containing those values. Proofs of these formulae can be found elsewhere (see 'Further study', p. 76).

Here we present both types of approach in solving regression problems. Remember to use the approach followed by your course or the approach you find the easiest.

Approach 1: Using original data (X, Y)

If we use the data given in the question as they stand (*original data*), then the following formulae will help us calculate m and c for the least squares line. These formulae make use of the sigma notation discussed in Chapter 2, p. 28.

Original data formulae

$$m = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

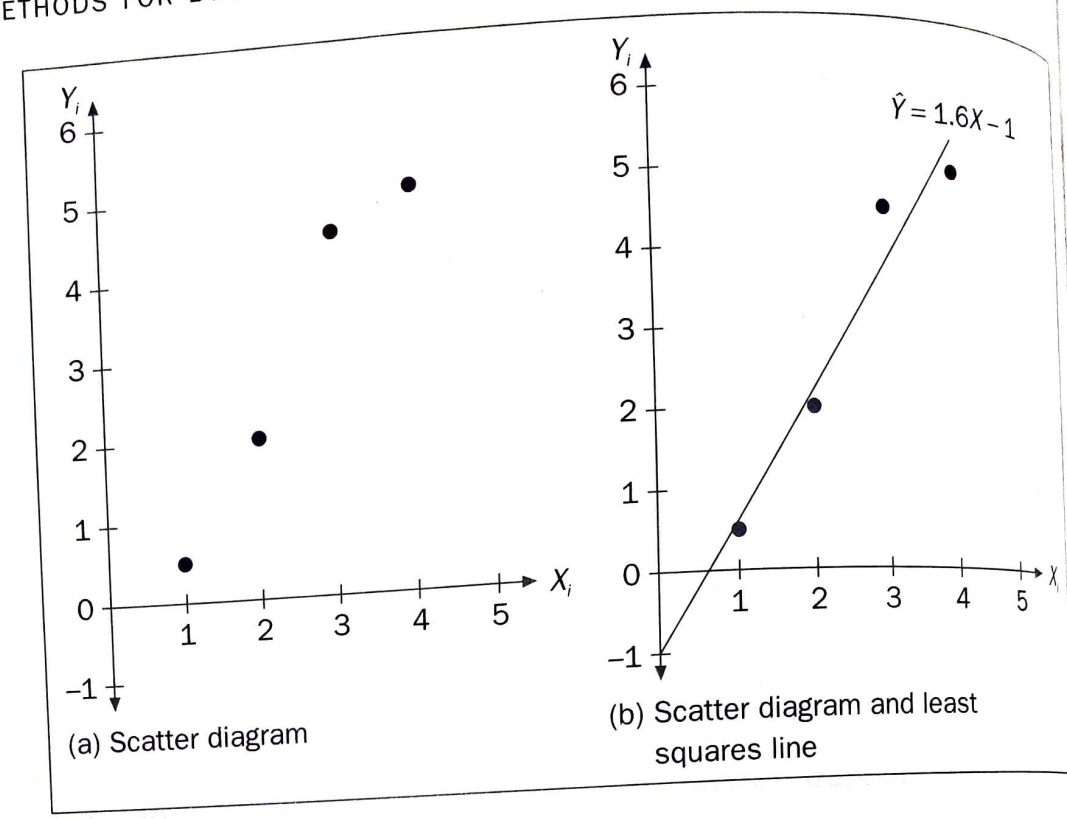
$$c = \bar{Y} - m\bar{X}$$

where n = number of observations

$\bar{X}, \bar{Y}$ = Arithmetic means of X_i, Y_i

WORKED EXAMPLE **3.1**

Suppose we wish to calculate the least squares line for the four observations shown on the scatter diagram (Figure 3.2(a)).



(a) Scatter diagram

(b) Scatter diagram and least squares line

figure 3.2
Using original data (X_i, Y_i) to calculate least squares line

Solution

We can set the data out in the columns needed to calculate m and c using the above formulae, as follows:

X_i	Y_i	X_iY_i	X_i^2
1	0.5	0.5	1
2	2	4	4
3	4.5	13.5	9
4	5	20	16
$\Sigma X = 10$	$\Sigma Y = 12$	$\Sigma XY = 38$	$\Sigma X^2 = 30$

Note: For simplicity we omit subscripts, etc. in the formulae.

$n = 4$ observations

$$\bar{X} = \frac{\sum X}{n} = \frac{10}{4} = 2.5$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{12}{4} = 3$$

$$m = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

$$m = \frac{4(38) - (10)(12)}{4(30) - (10)^2} = \frac{152 - 120}{120 - 100} = \frac{32}{20} = 1.6$$

$m = 1.6$

$$c = \bar{Y} - m\bar{X}$$

i.e. $c = 3 - 1.6\ (2.5)$

$c = 3 - 4$

$c = -1$

i.e. $\hat{Y} = 1.6X - 1$

Here the least squares line has a gradient (m) of +1.6, in other words the estimated value of Y is rising by 1.6 units for every 1 unit rise in X. The intercept (c) on the vertical axis is –1, in other words when X is 0 the estimated value of Y is –1. This least squares line is plotted in Figure 3.2(b).

Approach 2: Using redefined data

This approach can be missed out if your course uses Approach 1. In this case move directly to p. 57.

Rather simpler coding formulae can be derived using the fact that the regression (least squares) line must go through the point of means $(\bar{X},\bar{Y})$. For those interested, this fact is explained in selected sources in 'Further study' (p. 76). We can usefully redefine the origin from zero to the point of means $(\bar{X},\bar{Y})$, as in Figure 3.3, and use the resulting new axes (x_i, y_i) as the basis for our coding formulae. We can then calculate m at this new origin $(\bar{X},\bar{Y})$ more easily since the least squares line *goes through* the point of means $(\bar{X},\bar{Y})$ so that we no longer have any problem in calculating the vertical intercept at this origin. Of course we can then go back to the 'true' zero origin and find c once we have found m at the new origin $(\bar{X},\bar{Y})$.

If we use the *redefined data* for $x_i(X_i - \bar{X})$ and $y_i(Y_i - \bar{Y})$, then the following formulae will help us calculate m and c for the least squares line.

Coding formulae

Here we change the origin from zero to the point of means $(\bar{X},\bar{Y})$ and express new axes and data, $x_i\ (X_i - \bar{X})$ and $y_i\ (Y_i - \bar{Y})$

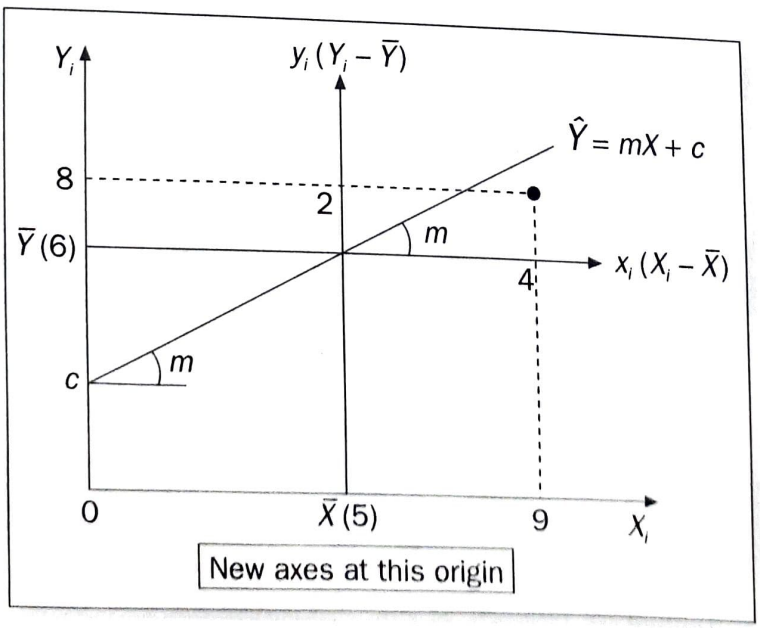


figure 3.3
Changing the origin from zero to the point of means $(\bar{X},\bar{Y})$ to establish the coding formulae. Here a single observation (9,8) is re-expressed as (4,2) using the new axes (x,y) when $\bar{X} = 5$ and $\bar{Y} = 6$.

$$m = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$c = \bar{Y} - m\bar{X}$$

Having calculated m (the slope of the least squares line) at the new origin $(\bar{X},\bar{Y})$ we revert to the zero origin to calculate c (the intercept). The slope, m, is of course identical at each of these origins for the linear least squares line.

Let us follow through our previous worked example, but this time using *redefined data*. The two approaches should give us the same result. Let us check that this is the case!

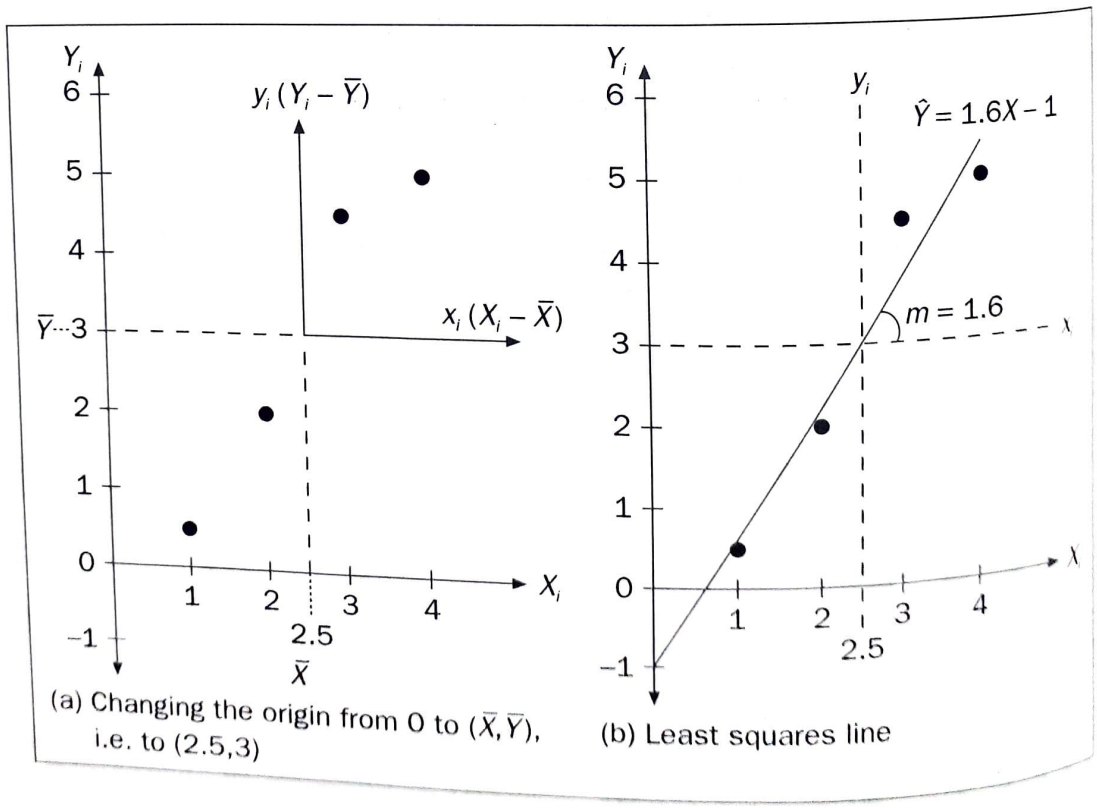
WORKED EXAMPLE **3.2**

Calculate the least squares line for the same four observations shown on the scatter diagram in Figure 3.2a (p. 54).

Solution

Here we use the new origin $(\bar{X},\bar{Y})$, namely (2.5,3) and *redefine* the data using the axes x_i $(X_i - 2.5)$ and y_i $(Y_i - 3)$.

figure 3.4
Using redefined data (x_i,y_i) to calculate the least squares line



(a) Changing the origin from 0 to $(\bar{X},\bar{Y})$, i.e. to (2.5,3)

(b) Least squares line

We can set the data out in the columns needed to calculate m and c using the coding formula for redefined data, as follows:

X_i	Y_i	x_i $(X_i - \bar{X})$	y_i $(Y_i - \bar{Y})$	$x_i y_i$	x_i^2
1	0.5	−1.5	−2.5	3.75	2.25
2	2	−0.5	−1	0.50	0.25
3	4.5	0.5	1.5	0.75	0.25
4	5	1.5	2	3.00	2.25
$\Sigma X = 10$	$\Sigma Y = 12$			$\Sigma xy = 8.00$	$\Sigma x^2 = 5.00$

$$\bar{X} = \frac{\sum X}{4} = \frac{10}{4} = 2.5$$

$$\bar{Y} = \frac{\sum Y}{4} = \frac{12}{4} = 3$$

$$m = \frac{\sum xy}{\sum x^2}$$

$$m = \frac{8.00}{5.00} = 1.6$$

$\underline{m = 1.6}$

We have now calculated m for the least squares line at the new origin, which is of course the same as m at the 'true' zero origin. We now revert to the zero origin to find c, the vertical intercept.

$$c = \bar{Y} - m\bar{X}$$

i.e. $c = 3 - 1.6(2.5)$

$$c = 3 - 4$$

$\underline{c = -1}$

i.e. $\underline{\hat{Y} = 1.6X - 1}$

We have the same result for the least squares line as in Approach 1, using the same data in each case. Clearly these two approaches are alternatives, yielding the same results.

DID YOU KNOW?

The founder of the 'least squares' regression technique was a French mathematician, Adrien-Marie Legendre. Born in 1752, he was awarded degrees in mathematics and physics. Having lost his wealth in the French Revolution in 1789 he was forced to 'work' for a living, lecturing in maths in Paris, working at the Greenwich Observatory in London and writing books. He developed the least squares technique in 1805, which was further developed by UK statistician Rober Gauss.

Prediction

Of course, once we have obtained our estimated least squares line we can use it for **prediction (forecasting)**. We can insert values of the independent variable (X_i) not yet experienced and predict values of the dependent variable (Y_i). Of course the confidence we can have in any such prediction will depend upon:

- the past relationship between Y and X continuing into the future,
- the estimated least squares line ($\hat{Y}$) fitting the past data rather well.

We return to the 'goodness of fit' of the least squares line in the next section. In our previous example we could substitute the value $X_i = 10$ into our equation:

$$\hat{Y} = 1.6X - 1$$
$$\hat{Y} = 1.6(10) - 1$$
$$\hat{Y} = 16 - 1 = 15$$

Hence our predicted value for Y is 15, should X be given the value of 10. We consider forecasting in more detail in Chapter 4.

PAUSE FOR THOUGHT 3.1 *Under what circumstances would you be reasonably confident in using your least squares line for prediction? Can you think of any problems that might arise to make these predictions unreliable?*

SELF-CHECK QUESTIONS

3.1 (a) Use the data below to create a scatter diagram.

X	1	2	3	4	5	6	7	8
Y	2	5	6	7	9	12	15	16

(b) Find the regression line relating *Y* (as dependent variable) to *X* (as independent variable).

(c) Comment on your results.

3.2 (a) Use the data below to create a scatter diagram.

X	1	2	3	4	5	6	7	8	9	10
Y	19	18	16	16	20	13	6	6	11	9

(b) Find the regression line relating *Y* (as dependent variable) to *X* (as independent variable).

(c) Comment on your results.

3.3 Find the 'line of best fit' (least squares line) for the following data.

X	4	5	6	12	13	13	16	16	17	17	18	19
Y	47	111	124	240	211	205	276	305	309	302	259	334
X	20	22	24									
Y	302	371	241									

Note: **Answers can be found on pp. 328–331.**

3.2 Correlation

The idea here is to measure how well the regression line fits the actual data. Two key measures are frequently used in this respect.

- the coefficient of determination (R^2)
- Pearson's coefficient of correlation (R)

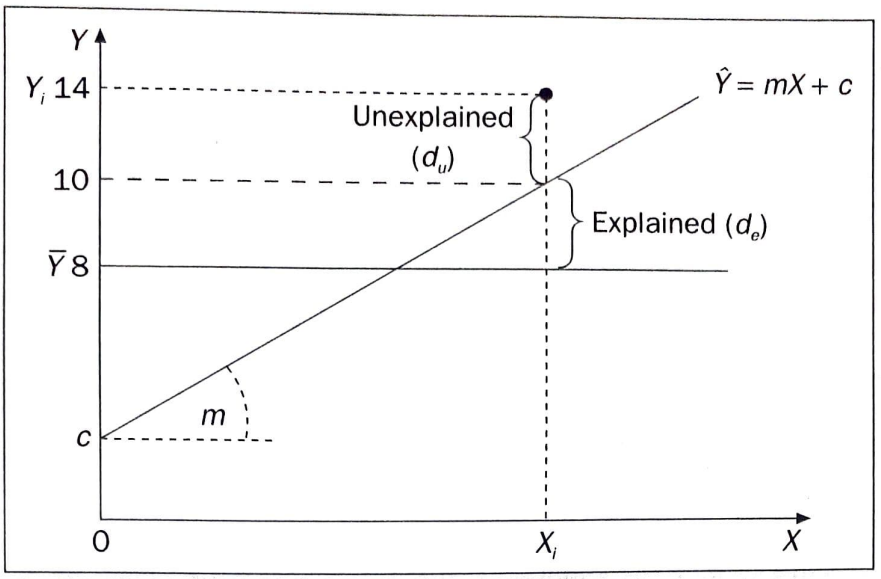


figure 3.5
Deviation for each observation (d_i) is the actual value (Y_i) minus the arithmetic mean ($\bar{Y}$). This can be split into two parts: explained deviation (d_e) and unexplained deviation (d_u)

Before considering *formulae* for calculating these measures, it will be helpful to define the concepts of deviation and variation which underpin them. Figure 3.5 will be used for illustration.

Deviation (d_i) is the difference between the actual value of an observation (Y_i) and its arithmetic mean ($\bar{Y}$). This deviation can be split into two separate parts. An explained part (d_e) which is predicted or accounted for by the regression line. An unexplained part (d_u) which is *not* predicted or accounted for by the regression line.

Deviation = $d_i = Y_i - \bar{Y}$
Deviation = explained deviation + unexplained deviation
Deviation = $d_i = d_e + d_u$

Summing such deviations across all n observations gives *total deviation.*

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} d_e + \sum_{i=1}^{n} d_u$$

i.e. total deviation = explained deviation + unexplained deviation

Variation (d_i^2) is the square of the difference between the actual value of an observation (Y_i) and its arithmetic mean ($\bar{Y}$). As before, squaring the deviations avoids the problem of first identifying, then dropping the sign ± of each deviation.

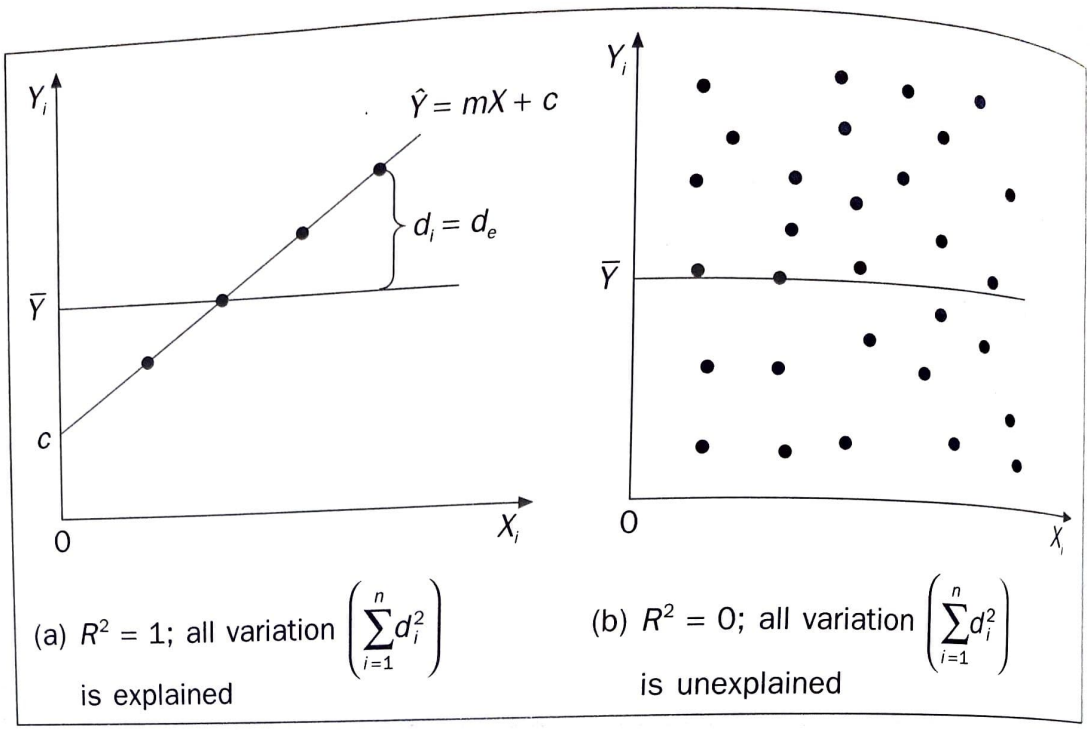
Summing the squared deviations across all n observations gives *total variation.*

It can be shown that:

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} d_e^2 + \sum_{i=1}^{n} d_u^2$$

i.e. total variation = explained variation + unexplained variation

figure 3.6
Coefficient of determination (R^2)



PAUSE FOR THOUGHT 3.2 *How does* total variation *differ from the* variance *discussed in Chapter 2?*

Coefficient of determination (R^2)

$$R^2 = \frac{\text{explained variation}}{\text{total variation}}$$

The **coefficient of determination** (R^2) is the ratio of 'explained variation' to 'total variation'. Clearly when $R^2 = 1$ as in Figure 3.6(a), then all the deviation and therefore variation can be explained or accounted for by the regression line. We have a perfect fit.

Similarly when $R^2 = 0$ (or close to 0) as in Figure 3.6(b), none of the deviation and therefore variation can be explained or accounted for by the regression line. We effectively have a random scatter of points to which any regression line fits as well as any other.

The closer R^2 is to 1, the better the fit of the least squares line to the actual data.

Pearson's coefficient of correlation (R)

$$R = \sqrt{\frac{\text{explained variation}}{\text{total variation}}}$$

Pearson's coefficient of correlation (R) is the square root of the coefficient of determination (R^2)

Just as R^2 varies between 1 and 0, so R varies between ±1 and 0. 'Perfect correlation' implies a value for R of +1 when the relationship between Y and X is direct, and −1 when the relationship between Y and X is indirect (inverse).

figure 3.7
Coefficient of correlation (*R*)

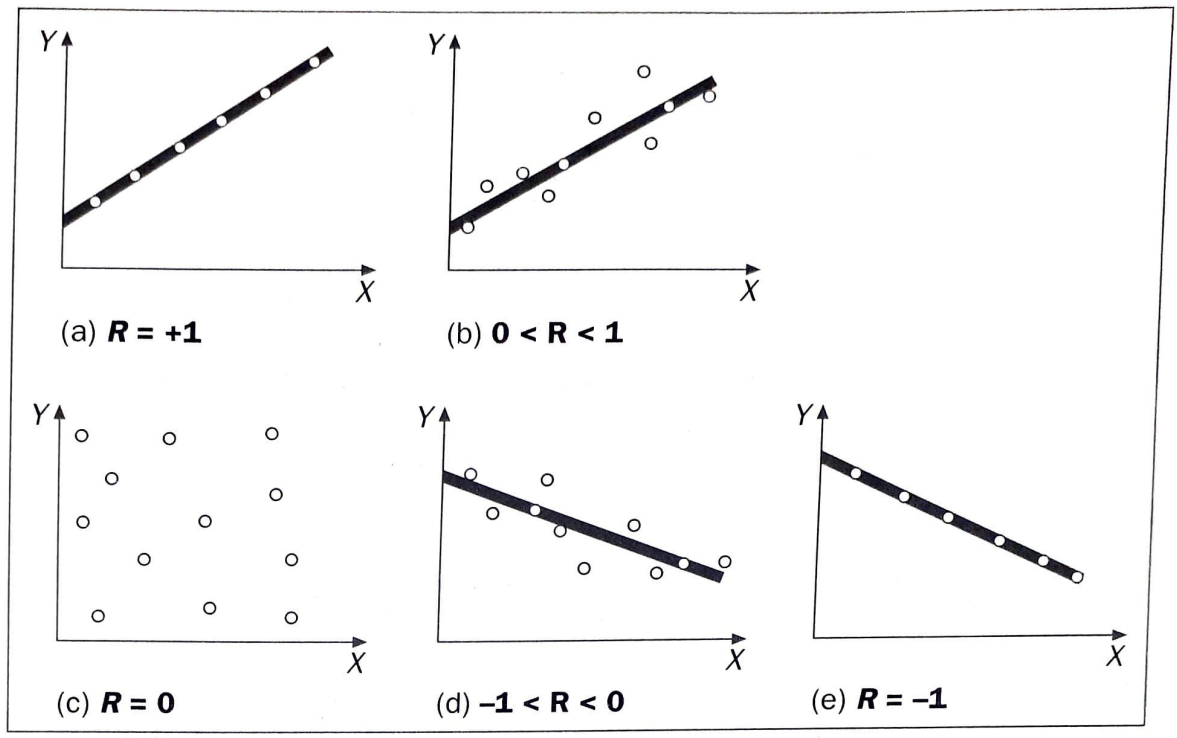


Figure 3.7 outlines the various possible values for R, the coefficient of correlation. This is often called Pearson's coefficient of correlation and can have values between +1 and –1.

(a) ***R* = +1**
This is perfect positive correlation. Here the relationship between Y and X is direct (positive), as both variables rise or fall together. The sign for R is therefore positive, and the value is +1 since the fit of the least squares line is perfect (all the variation is explained).

(b) ***R* > 0 but < 1**
Again a direct (positive) relationship between Y and X, so the sign is positive for R. Value for R is < 1, since not all of the variation is explained.

(c) ***R* = 0 (random scatter)**
Essentially no relationship between Y and X. All the variation is unexplained.

(d) ***R* < 0 but > –1**
Here the relationship between Y and X is indirect (negative or inverse) as each variable moves in the opposite direction; when one variable rises the other falls, and vice verse. The sign for R is therefore negative but the value is greater than –1 since not all of the variation is explained.

(e) ***R* = –1**
This is perfect negative correlation. The relationship between Y and X is indirect (negative or inverse) and the value is –1 since the fit of the least squares line is perfect (all the variation is explained).

DID YOU KNOW?

The easiest way to interpret the meaning of a particular value of R, the Pearson coefficient of correlation, is to square it. For example a value $R = +0.8$ means that $R^2 = (+0.8)^2 = 0.64$, i.e. 64 per cent of total variation can be 'explained' or accounted for by your regression line.

Calculating R^2 and R

As when calculating the least squares line, two alternative approaches can be used to derive formulae for calculating R^2 and R. Be consistent, and follow the same approach here as you have done for calculating the regression line.

Approach 1: Using original data (X,Y)

Finding R^2 and R using original data

$$R^2 = \left[\frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right] \times \left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i \right)^2 \right]}} \right]^2$$

where n = number of observations

$R = \sqrt{R^2}$

WORKED EXAMPLE **3.3**

Here we apply our formula to the data and calculations already presented on p. 54. We reproduce this below, noting that we need an *extra column* for Y_i^2.

X_i	Y_i	X_iY_i	X_i^2	Y_i^2
1	0.5	0.5	1	0.25
2	2	4	4	4
3	4.5	13.5	9	20.25
4	5	20	16	25
$\Sigma X = 10$	$\Sigma Y = 12$	$\Sigma XY = 38$	$\Sigma X^2 = 30$	$\Sigma Y^2 = 49.50$

$$R^2 = \left[\frac{4(38) - (10)(12)}{\sqrt{\left[4(30) - (10)^2\right] \times \left[4(49.5) - (12)^2\right]}} \right]^2$$

$$R^2 = \left[\frac{152 - 120}{\sqrt{\left[120 - 100\right] \times \left[198 - 144\right]}} \right]^2$$

$$R^2 = \left[\frac{32}{\sqrt{20 \times 54}} \right]^2 = \left[\frac{32}{\sqrt{1080}} \right]^2 = \left[\frac{32}{32.86} \right]^2$$

$R^2 = 0.948$

In other words 0.948 or, as a percentage, 94.8 per cent of the total variation is 'explained' or accounted for by our regression line. This indicates a very good fit of our least squares line to the data, giving us greater confidence should we use our regression line for prediction (forecasting).

Approach 2: Using redefined data

Alternatively, we can transfer the origin from zero to the point of means $(\bar{X},\bar{Y})$ and use a coding formula for R^2 and R. In this case we make use of new axes $x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$.

Finding R^2 and R using the coding formula

Here we can make use of calculations already undertaken in finding m and c for the regression line $Y = mX + c$ when using the coding formula

$$R^2 = \frac{\left[\sum_{i=1}^{n} x_i y_i\right]^2}{\left[\sum_{i=1}^{n} x_i^2\right]\left[\sum_{i=1}^{n} y_i^2\right]}$$

where $x_i = X_i - \bar{X}$

$y_i = Y_i - \bar{Y}$

$R = \sqrt{R^2}$

WORKED EXAMPLE **3.4**

Here we apply our coding formula to the data and calculations already presented on p. 57. We reproduce this below, noting that this time we need an *extra column* for y_i^2.

Solution

X_i	Y_i	x_i $(X_i - \bar{X})$	y_i $(Y_i - \bar{Y})$	$x_i y_i$	x_i^2	y_i^2
1	0.5	−1.5	−2.5	3.75	2.25	6.25
2	2	−0.5	−1	0.50	0.25	1
3	4.5	0.5	1.5	0.75	0.25	2.25
4	5	1.5	2	3.00	2.25	4
$\Sigma X = 10$	$\Sigma Y = 12$			$\Sigma xy = 8.00$	$\Sigma x^2 = 5.00$	$\Sigma y^2 = 13.50$

$$\bar{X} = \frac{\Sigma X}{4} = \frac{10}{4} = 2.5$$

$$\bar{Y} = \frac{\Sigma Y}{4} = \frac{12}{4} = 3$$

Using our coding formula

$$R^2 = \frac{[\sum xy]^2}{[\sum x^2][\sum y^2]} = \frac{(8.00)^2}{(5.00)(13.50)} = \frac{64.00}{67.50} = 0.948$$

$\underline{R^2 = 0.948}$

Again, note that either approach will give the same solution.

WORKED EXAMPLE **3.5**

The following worked example uses *both* approaches to calculate the regression line and the coefficients of determination and correlation for a data set containing six observations.

Approach 1 solution: Using original data

X_i	Y_i	X_iY_i	X_i^2	Y_i^2
0	−4	0	0	16
1	−3	−3	1	9
2	−1	−2	4	1
3	2	6	9	4
4	3	12	16	9
5	6	30	25	36
$\Sigma X = 15$	$\Sigma Y = 3$	$\Sigma XY = 43$	$\Sigma X^2 = 55$	$\Sigma Y^2 = 75$

$$m = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$m = \frac{6(43) - (15)(3)}{6(55) - (15)^2} = \frac{258 - 45}{330 - 225} = \frac{213}{105}$$

$\underline{m = +2.03}$

$$c = \bar{Y} - m\bar{X}$$

$\bar{Y} = 0.5, \quad \bar{X} = 2.5$

$c = 0.5 - 2.03(2.5)$

$\underline{c = -4.6}$

$$\underline{\hat{Y} = 2.03X - 4.6}$$

$$R^2 = \left[\frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right] \times \left[n\sum Y^2 - (\sum Y)^2\right]}}\right]^2$$

$$R^2 = \left[\frac{6(43) - (15)(3)}{\sqrt{\left[6(55) - (15)^2\right] \times \left[6(75) - (3)^2\right]}}\right]^2$$

$$R^2 = \left[\frac{258 - 45}{\sqrt{[105] \times [441]}}\right]^2$$

$$R^2 = \left[\frac{213}{\sqrt{46{,}305}}\right]^2$$

$$R^2 = \left[\frac{213}{215.2}\right]^2$$

$$\underline{R^2 = 0.98}$$

$$R = \sqrt{0.98}$$

$$\underline{R = 0.99}$$

PAUSE FOR THOUGHT 3.3 *Plot the scatter diagram from the data for this Worked example. Now impose the least squares line on your scatter diagram. Would you have expected to find a high R^2?*

Approach 2 solution: Using redefined data

X_i	Y_i	x_i $(X_i - \bar{X})$	y_i $(Y_i - \bar{Y})$	$x_i y_i$	x_i^2	y_i^2
0	−4	−2.5	−4.5	11.25	6.25	20.25
1	−3	−1.5	−3.5	5.25	2.25	12.25
2	−1	−0.5	−1.5	0.75	0.25	2.25
3	2	0.5	1.5	0.75	0.25	2.25
4	3	1.5	2.5	3.75	2.25	6.25
5	6	2.5	5.5	13.75	6.25	30.25
$\Sigma X = 15$	$\Sigma Y = 3$			$\Sigma xy = 35.5$	$\Sigma x^2 = 17.50$	$\Sigma y^2 = 73.50$
$\bar{X} = 2.5$	$\bar{Y} = 0.5$					

$$m = \frac{\sum xy}{\sum x^2} \qquad c = \bar{Y} - m\bar{X} \qquad R^2 = \frac{\left(\sum x_i y_i\right)^2}{\left(\sum x_i^2\right)\left(\sum y_i^2\right)}$$

$$m = \frac{35.5}{17.5} \qquad c = 0.5 - (2.03)2.5 \qquad R^2 = \frac{1{,}260.25}{1{,}286.25}$$

$$R^2 = 0.98$$

$$\underline{m = 2.03} \qquad \underline{c = -4.6}$$

$$R = \sqrt{0.98}$$

$$\underline{\hat{Y} = 2.03X - 4.6} \qquad \underline{R = 0.99}$$

ACTIVITY **3**

The intention here is to derive a regression (least squares) line relating the population of Greece (Y) to time (X). This is important to the long-term planning of our company which sees Greece as a major export market.

You can use the data of Activity 2 (Chapter 2, pp. 45–46) and a spreadsheet with which you are familiar. The approach used involves the coding formula (see p. 55). You can solve either from first principles or use the in-built functions of your spreadsheet.

Solving from first principles

1 Set up your worksheet as follows:
Either load your original data of the Greek population – from Activity 2 (Chapter 2, pp. 45–46) and copy the data from your original sheet to a new clean sheet or re-enter the data into a new spreadsheet. You should end up with something similar to Figure A3.1 but without the formulae.

figure A3.1

	A	B	C	D	E	F	G	H	I
1									
2			**Xbar**	**Ybar**		**m (slope)**	**c (intercept)**		**R^2**
3			=AVERAGE(C6:C19)	→		=G20/H20	=D3–(F3*C3)		=G20^2/(H20*I20)
4			**X**	**Y**					
5		**y hat**	**year**	**population**	**xi**	**yi**	**xiyi**	**xi^2**	**yi^2**
6		=$G$3+($F$3*C6)	1986	9,642,505	=C6–$C$3	=D6–$D$3	=E6*F6	=E6^2	=F6^2
7			1987	9,729,350					
8			1988	9,789,513					
9			1989	9,846,627					
10			1990	9,895,801					
11			1991	9,934,249					
12			1992	9,963,604					
13			1993	9,983,490					
14			1994	10,004,401					
15			1995	10,038,672					
16			1996	10,088,700					
17			1997	10,200,000					
18			1998	10,311,300					
19			1999	10,422,600	↓	↓	↓	↓	↓
20			2000						
21			2001			**Sums**	=SUM(G6:G19)	→	→
22		↓	2002						

2 Find the value of m (slope)

$$m = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

(a) In cells **C3 & D3** find the mean of the X and Y values.
(b) In cell **E6** calculate x_i – where $x_i = X_i - \bar{X}$.
(c) In cell **F6** calculate y_i where $y_i = Y_i - \bar{Y}$.
(d) In cell **G6** multiply x_i by y_i.
(e) In cell **H6** square x_i.

(f) Highlight cells **E6 to H6** and copy the formulae down to **Row 19**.
(g) In cell **G20 & H20** find the sums of $x_i y_i$ and x_i^2.
(h) In cell **F3** calculate m by dividing the sum of $x_i y_i$ by the sum of x_i^2.

3 ***Find the value of c (intercept) where:*** $c = \bar{Y} - m\bar{X}$.

Your table already contains all the values needed to substitute in the above equation, therefore, enter the formula in Figure A3.1 cell G3 to calculate the intercept.

4 ***Calculate the coefficient of determination*** R^2.

Use the equation:

$$R^2 = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)}$$

Here we can make use of the calculations already undertaken in finding m and c for the regression line.

(a) Extend your table by adding the heading *yi^2* in cell **I5**.
(b) In cell **I6** calculate y_i^2.
(c) Copy the formula down to cell **I19**.
(d) In cell **I20** find the sum of y_i^2.
(e) In cell **I3** enter the formula for R^2.

5 ***Forecast the values of X using the equation:*** $\hat{Y} = c + mX$.

(a) Add the heading **yhat** in cell **B5** and the years to be predicted in cells **C20 to C22**.
(b) In cell **B6** calculate $\hat{Y}$ for $X = 1986$.
(c) Copy the formulae down to cell **B22**.

Cells **B6 to B22** hold the predicted values from the equation:

$$\hat{Y} = -90037801.5 + 50277.5(X)$$

B20 to B22 hold the predictions for the years 2000–2002.
Figure A3.3 (on p. 465) shows the complete solution.

Using in-built functions

1 ***Select a new unused worksheet and enter the following labels:***

	B	C	D
2	Slope	Intercept	R^2
3			
4			
5	Y	X	
6	Population	Year	Y hat

Copy the dependent data (Y) – the Greek population data into cells **B7 to B20**.
Enter the independent data (X) 1986–2002 into cells **C7 to C23**.
You should end up with something similar to Figure A3.2 below.

figure A3.2

	A	B	C	D
1				
2		**slope**	**intercept**	**R^2**
3		=SLOPE(B7:B20,C7:C20)	=INTERCEPT(B7:B20,C7:C20)	=RSQ(B7:B20,C7:C20)
4				
5		**Y**	**X**	
6		**population**	**year**	**Y hat**
7		9,642,505	1986	=$B$3*C7+$C$3
8		9,729,350	1987	
9		9,789,513	1988	
10		9,846,627	1989	
11		9,895,801	1990	
12		9,934,249	1991	
13		9,963,604	1992	
14		9,983,490	1993	
15		10,004,401	1994	
16		10,038,672	1995	
17		10,088,700	1996	
18		10,200,000	1997	
19		10,311,300	1998	
20		10,422,600	1999	
21			2000	
22			2001	
23			2002	

2 Calculate m, c and R^2.

In cells **B3:D3** use the in-built functions to calculate the slope or gradient, intercept and R^2. In *Microsoft's Excel*, for example, the following functions and their arguments are used:

= *Slope (known Ys,known Xs)*

= *Intercept (known Ys,known Xs)*

= RSQ *(known Ys,known Xs)*

3 Calculate $\hat{Y}$ where $\hat{Y} = c + mX$

(a) Add the years to be predicted (2000, 2001, 2002) into cells **B21 to B23**

(b) In cell **D7** substitute the values for *m and c* into the equation

(c) Copy the formulae down to cell **B23**

Cells **D7 to D23** hold the predicted values from the equation:

$$\hat{Y} = -90037801.5 + 50277.5(X)$$

Thus cells **D21 to D23** hold the predicted values for the years 2000–2002. Figure A3.4 (p. 465) shows the complete solution.

3.3 Spearman's coefficient of rank correlation

Sometimes data are presented in *rank order*, whether descending (first to last) or ascending (last to first). It may be that the ranking comes from two separate

sources, as when applicants for a post are ranked both by interview and by psychological tests. We might then wish to check to what extent the different rankings agree!

In such situations **Spearman's coefficient of rank correlation** (R_s) might be calculated. The formula for R_s is as follows:

$$R_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

where R_s = Spearman's rank correlation coefficient
d_i = difference in ranking of a given observation
n = number of observations

If $R_s = 1$, then clearly we have perfect agreement between the respective rankings. This follows from the fact that each d_i would be zero when the respective rankings are identical, giving $R_s = 1 - 0$ in the above expression. The closer to 1 the value of R_s, the better the agreement between the respective rankings.

WORKED EXAMPLE **3.6**

Ten candidates for a managerial post were ranked by interview and psychological test in the following manner:

	Candidates rankings									
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
I Interview	4	2	7	1	5	6	9	3	10	8
II Psychological test	3	2	5	1	4	9	6	7	8	10

Calculate Spearman's rank correlation coefficient and discuss whether it represents a measure of agreement between the two types of test.

Solution

The first step is to take the difference in the ranks (d_i); the second step is to square each of these differences; the third step is to sum the squares of the differences. We then substitute the result into the formula

d_i	1	0	2	0	1	3	3	4	2	2
d_i^2	1	0	4	0	1	9	9	16	4	4

$$\sum_{i=1}^{10} d_i^2 = 48$$

$$R_s = 1 - \frac{6\sum d_i^2}{n^3 - n} = 1 - \frac{6 \times 48}{10^3 - 10} = 1 - \frac{288}{990} = 0.709$$

The Spearman's rank correlation coefficient shows *some* agreement in ranking between the two types of procedure.

We can use Spearman's coefficient of rank correlation (R_s) whenever the data for two variables are given in *rank order* (i.e. the data are **ordinal**). Whenever data are given in terms of *absolute values* (i.e. the data are **cardinal**) we would usually use the earlier coefficient of correlation (R). However we can *convert* two sets of data which are given in absolute values into rank order should we wish to calculate R_s. Further, should one set of data be given in terms of absolute values (cardinal) and the other in terms of rank order (ordinal) then only by converting the cardinal data to ordinal and calculating R_s might we be able to establish a measure of 'goodness of fit' between the two types of data.

SELF-CHECK QUESTIONS

3.4 Using the data for Self-check question 3.1 (p. 58)

(a) Find the coefficient of determination.
(b) Find Pearson's coefficient of correlation.
(c) Comment on your results.

3.5 Using the data for Self-check question 3.2 (p. 58)

(a) Find the coefficient of determination.
(b) Find Pearson's coefficient of correlation.
(c) Comment on your results.

3.6 The data below give the actual sales of a company (£000s) in each of eight regions of a country together with the forecast of sales by two different methods.

Region	*Actual sales (£000)*	*Forecast: Method 1*	*Forecast: Method 2*
A	15	13	16
B	19	25	19
C	30	23	26
D	12	26	14
E	58	48	65
F	10	15	19
G	23	28	27
H	17	10	22

(i) Calculate the rank correlation coefficient between
(1) Actual sales and forecast 1
(2) Actual sales and forecast 2

(ii) Which forecasting method would you recommend next year?

Note: **Answers can be found on pp. 331–334.**

3.4 Multiple regression

Our analysis has, until this point, concentrated on relationships involving only two variables: a dependent variable Y and an independent variable X. In reality

many relationships involve more than two variables, and attempts to fit lines (or surfaces) of 'best fit' to such variables is termed **multiple regression** analysis. Here we merely touch upon a few of the issues involved, with more detailed sources of information indicated on p. 76.

Win Hornby *et al.* (2001) provide a useful illustration of the application of multiple regression analysis in their review of a study seeking to relate the demand for beer in the UK to as many as *seven* separate independent variables, namely:

1 P_b the real price of beer (i.e. the price of beer in relation to the rate of inflation).
2 A_b the real level of advertising on beer (i.e. the level of advertising in relation to an advertising index of TV and press).
3 A_a the real level of advertising on other alcoholic drinks.
4 W the daily mean temperature.
5 Y_r the real level of personal disposable income.
6 P_c the real price of other alcoholic drink.
7 S a seasonal factor to reflect the higher consumption of beer in the summer months and at Christmas.

Clearly we no longer have a two-dimensional scatter diagram. Instead of a line of 'best fit' we now have a plane or surface which 'best fits' a multi-dimensional scatter diagram. Figure 11.9 (p. 272) gives an idea of a 'surface' in a three-dimensional situation. Statistical packages are frequently used to calculate such relationships.

Using quarterly data stretching back over a 15 year period the authors of this study, W. D. Reekie and C. Blight, established the following results from 60 data observations:

$$Q = 3.18 - 1.355P_b + 0.004A_b + 0.019A_a + 0.057W + 0.085Y_r + 0.371P_c + 0.0195S$$
$$\quad (0.287) \quad (0.64) \quad (0.32) \quad (2.66)^* \quad (5.84)^* \quad (0.10) \quad (0.173)$$

where Q = demand for beer in the UK (millions of barrels per year)

$R^2 = 0.57$

t-ratios are in brackets. * indicates the statistically significant variables.

- Notice that the *signs* in front of six of the independent variables are *positive*, suggesting that a rise in those variables will *increase* the demand for beer in the UK (shift the demand curve to the right).
- However the sign in front of P_b, the real price of beer, is negative, suggesting that a rise in the price of beer will result in a *fall* in the demand for beer (i.e. a contraction in demand, moving up the demand curve from right to left).
- The numbers (coefficients) in front of each independent variable indicate the *magnitude* of the respective relationships. For example, any given

percentage change in the price of beer will have the largest (here negative) impact on the demand for beer as compared to changes in the other independent variables.

- The large 'intercept' term of 3.18 indicates that should these seven independent variables have zero values, there would still be a sizeable demand for beer in the UK (3.18 million barrels per year).
- The R^2 figures of 0.57 indicates that this specified relationship only explains (accounts for) some 57 per cent of the total variation (see p. 60 above). This means that some 43 per cent of the total variation is unexplained. For example we may have omitted one or more independent variables which really do have an impact on the demand for beer in the UK.
- The figures in brackets refer to the t statistic scores (or t ratios) discussed in Chapters 6 and 7. The null hypothesis (see p. 159) for the t test conducted on each independent variable is that the coefficient in front of that independent variable is zero. The larger the t score, the more confident we can be in *rejecting* the null hypothesis, i.e. accepting the alternative hypothesis, namely that the independent variable in question really does have a non-zero coefficient. Put another way, the larger the t score, the more confident we can be in concluding that the particular independent variable is 'significant' in influencing the dependent variable. In the results presented above, we can be most confident about the influence on the demand for beer of expenditure on advertising on other alcoholic drinks and the daily mean temperature.

Note: Perhaps surprisingly the sign of the coefficient in front of A_a is positive; suggesting that as more is spent on advertising *other* alcoholic drinks, the demand for beer increases. This suggests positive 'spillover' effects, namely that alcoholic drinks are a generic product, and as one type is advertised so other types of alcoholic drinks also benefit.

REVIEW QUESTIONS

3.1 Sales of products for the past 12 months are shown below. Use linear regression to forecast sales for the next three months.

Month	*Jan.*	*Feb.*	*March*	*April*	*May*	*June*	*July*	*Aug.*	*Sept.*	*Oct.*	*Nov.*	*Dec.*
Sales	89	74	92	123	102	139	152	128	183	148	166	179

Comment on the reliability of your prediction.

3.2 The weather was considered an important element in the sales pattern of a certain product. Over a four month period the average daily air temperature was recorded and the average daily sales of the product were calculated for each temperature. The results were as follows:

Temp (°C)	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*
Average daily sales	103	104	95	83	81	75	68	47	53	40	43	40	38	39	35

(a) Plot a scatter diagram of the data. What does this suggest?
(b) Calculate the linear regression equation that best fits the data.
(c) Plot the least squares regression line on your scatter diagram.
(d) Use the regression equation to estimate the number of products that will be sold if the temperature reaches 25°C.

3.3 The data below show the lateness of trains (in minutes) arriving at a terminus station and the number of passengers alighting from those trains. The train operator believes that the number of passengers (Y) will be greater if the train arrives a little later (X) than scheduled.

Minutes late	***No. of passengers***
1	1,459
3	395
0	534
2	641
7	927
3	650
4	447
5	392
0	569
11	713
1	401
12	2,691
1	443
7	883
12	1,577
11	147
7	568
2	455
3	632
4	531

(a) Plot a scatter diagram of the data.
(b) Calculate the linear regression equation that best fits the data.
(c) Plot the least squares regression line on your scatter diagram.
(d) Is their any evidence to suggest that there is a relationship between the number of passengers and the punctuality of the train.

3.4 A manager needs to determine the relationship between the company's advertising expenditure and its sales revenue. If the relationship is a strong one, then the company will be able to forecast the sales revenue for a given advertising expenditure. The following data have been collected over the past eight years.

Advertising (£000s)	*Sales (£000s)*
8.73	40.285
8.78	39.423
9.25	47.389
9.54	45.707
10.29	47.711
11.22	52.500
12.92	53.727
13.99	64.389

(a) Plot a scatter diagram of the data.
(b) Calculate the linear regression equation that best fits the data.
(c) Estimate the sales revenue when advertising is:
£11,000
£13,500
£15,000
(d) Is the relationship sufficiently strong to use as a predictive method?

3.5 The data below show a hypothetical index of economic activity and the sales revenue for a firm over a 14 year period.

Index of economic activity	*Sales revenue (£000)*
103	1,188
104	1,224
104	1,243
105	1,239
105	1,242
107	1,138
108	1,481
108	1,562
109	1,554
109	1,566
110	1,598
111	1,562
112	1,609
112	1,499

(a) Calculate the regression line relating the two sets of data. Let sales revenue be the dependent variable. Comment on your result.
(b) How well does your regression line fit the data.

3.6 A quality control manager believes that the number of rejected items on an assembly line is related to the length of time between breaks (the work period) for the assembly staff. The period between breaks is measured as well as the percentage of rejects produced by the operatives with the following results:

Work period (minutes)	Rejects %
60	2.3
65	2.5
70	2.7
75	2.9
80	3.2
85	3.6
90	3.5
95	3.7
100	4.0
105	4.4
110	4.3
115	4.6
120	4.7
125	3.3
130	5.1
135	5.2

(a) Calculate the regression line and use it to consider if there is any evidence to support the quality control manager's theory.
(b) Given the above data, is there a case to suggest that work periods not exceeding two hours would reduce the number of rejects below 5 per cent?
(c) How well does your regression line fit the data?

3.7 The marketing department believes that consumer tastes for beer are related to geographical area. They have therefore commissioned research to investigate the popularity of six leading brands of beer. Two different panels, one from the north the other from the south, tasted the products and ranked them in their order of preference. To avoid bias, the name of the products was withheld and substituted with the labels A–F. The results of the blind tasting were as follows:

Product	Northern panel	Southern panel
A	3	2
B	4	4
C	5	3
D	2	6
E	6	5
F	1	1

Is there any statistical evidence to suggest that the tastes of consumers are related to geographical region?

3.8 A motoring organisation commissioned two independent testing laboratories to assess the safety of eight of the best-selling models of car in the mid-range family saloon market. The laboratories used different testing procedures and methods before ranking the models in order of safety (where 1 = safest, 8 = least safe) as follows:

Product	Lab 1	Lab 2
Chimera 205	1	2
Apollo T7	6	5
Pastiche	7	8
Ganymede	4	6
Callisto LX	3	4
QM4	5	3
Alfredo GTi	8	7
Arctura	2	1

Is there any correlation between the two sets of results?

Note: Answers can be found on pp. 392–402.

Further study and data

Texts

Bancroft, G. and O'Sullivan, G. (1993), *Quantitative methods for accounting and business studies*, 3rd edn, McGraw Hill, chapter 9.

Curwin, J. and Slater, R. (1996), *Quantitative methods for business decisions*, 4th edn, International Thompson Business Press, chapters 17, 18 and 19.

Hornby, W., Gammie, B. and Wall, S. (2001), *Business Economics*, 2nd edn, Financial Times Prentice Hall, chapter 6.

Lawson, M., Hubbard, S. and Pugh, P. (1995), *Maths and statistics for business*, Addison Wesley Longman.

Morris, C. (1999), *Quantitative approaches in business studies*, 5th edn, Pitman, chapters 13, 14 and 15.

Oakshott, L. (1996), *Essential elements of business statistics*, BPP, chapter 8.

Swift, L. (2001), *Quantitatve methods for business, management and finance*, Palgrave Publishers, part S3.

Thomas, R. (1997), *Quantitative methods for business studies*, Prentice Hall, chapter 3.

Waters, D. (1997), *Quantitative methods for business*, 2nd edn, Addison Wesley Longman, chapter 8.

Wisniewski, M. with Stead, R. (1996), *Foundation quantitative methods for business*, Pitman, chapter 12.

Sources of information and data

See list at end of Chapter 1 (p. 25).

On-line exercises

Check the web site *www.booksites.net/burton* to find extra questions, Spreadsheet and on-line exercises for this chapter.

CHAPTER 4

Time series and forecasting

Objectives

By the end of this chapter you should be able to:

- use time series data to find both trend and seasonal variation as a basis for forecasting;
- apply the technique of moving averages to help in the calculation of trend lines and seasonal factors;
- combine time series and regression analysis for purposes of forecasting;
- apply exponential smoothing as a forecasting technique;
- be aware of the use of other approaches to forecasting in contemporary business practice.

Introduction

Data can be collected for purposes of analysis in a number of ways. For example data for household expenditure on food can be collected in different locations (nation, region, city, etc.) at a *single point of time*. These are called **cross-sectional data**. Alternatively data for household expenditure on food can be collected for a given location (nation, region, city, etc.) at *different points in time*. This is called **time series data**.

In this chapter we consider the use of such **time series** data as a basis for estimating both trend and seasonal variation. As we shall see, such estimates can then be used as a basis for **forecasting**.

In fact we have already touched on forecasting in Chapter 3 when we used the least squares line to predict values of the dependent variable (Y) for as yet unobtained values of the independent variable (X). We take this aspect of forecasting using regression analysis a little further in this chapter. Indeed we see how we can usefully *combine* regression analysis with the results of our time series analysis to forecast future outcomes.

The chapter concludes with a more general review of forecasting techniques, including **exponential smoothing** and other widely used techniques in current business practice.

Answers to the 'Self-check questions', responses to each 'Pause for thought' and answers to the 'Review questions' can be found at the end of the book.

4.1 Time series analysis

Many types of data have been recorded through time, such as monthly, quarterly or annual data for output, sales revenue, profit, employment, unemployment, prices, etc.

Components of time series

Typically such **time series** data exhibit one or more of the following components.

- A **trend component (*T*)**, whereby the variable appears to broadly rise or fall (or remain unchanged) through time.
- A **seasonal component (*S*)**, whereby the variable moves in regular cycles *within a year* around that trend line. For example within a broadly rising value of sales through time (trend), an ice cream manufacturer may regularly experience peaks for sales in summer months and troughs for sales in winter months (seasonal).

Note that although the term 'seasonal' has been applied to such short-term cycles, these may involve any units of time within a year; for example days within a week (e.g. high sales on Fridays and low sales on Mondays), weeks within a month, months or quarters or 'seasons' within a year.

PAUSE FOR THOUGHT 4.1 *Can you list six products (goods or services) other than the ice cream already mentioned, the demand for which you might expect to experience seasonal variation within a year?*

- A **cyclical component** (**C**), whereby the variable moves in a rather less regular cycle over the medium to longer term around the trend line. For example some have claimed to observe a *business cycle* of some eight to ten years between periods of 'boom' and 'bust' in modern industrialised economies. Others have claimed to observe still longer cycles over 50 years, with the peaks of such cycles related to new technological breakthroughs, such as water power, steam, electricity, microelectronics, etc. These have sometimes been referred to as *Kondratief cycles*, after a Russian economist of that name.

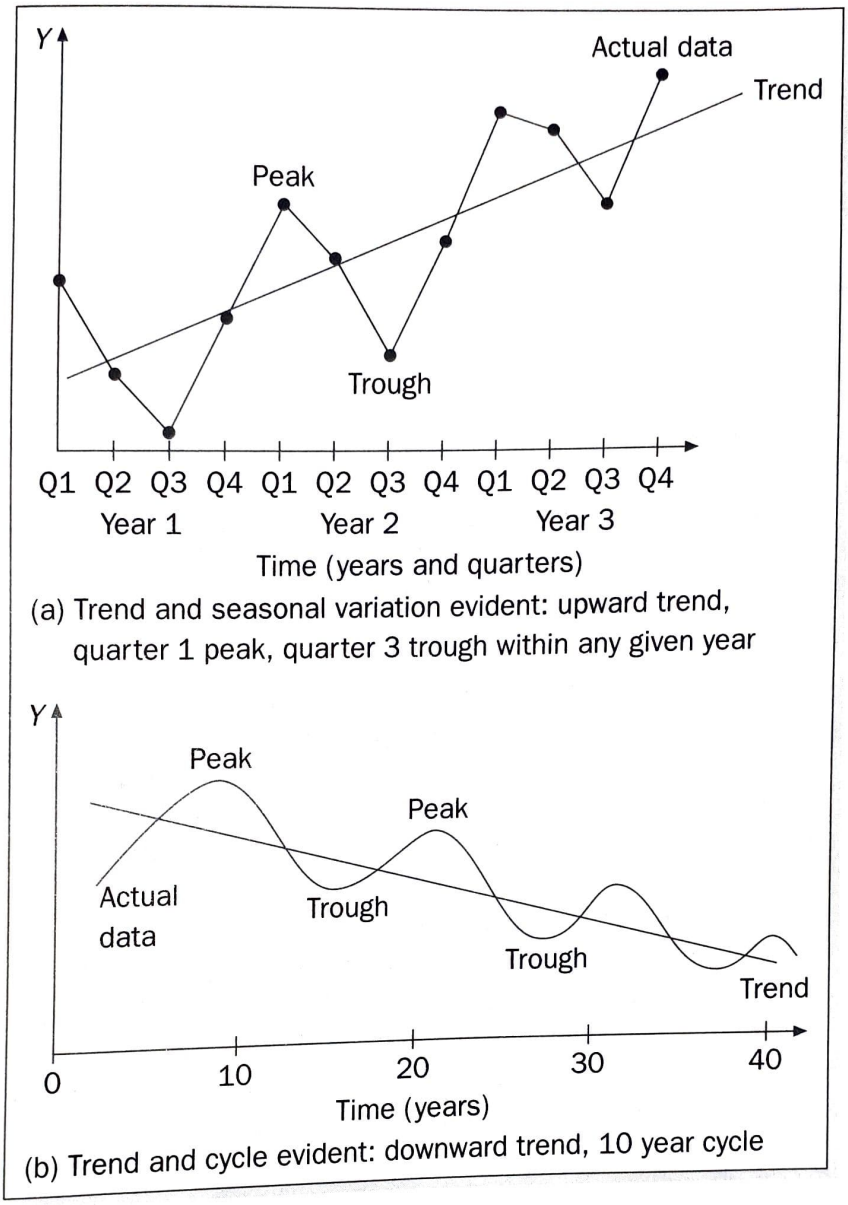
Note that the **period** of a cycle is often referred to as the time between successive peaks or successive troughs.

- An **irregular component** (**I**), which is entirely unpredictable. The stock market crash of 1987 was, for example, a dramatic and unexpected departure from trend in terms of share prices.

PAUSE FOR THOUGHT 4.2 *Can you name three other irregular components which have influenced a wide range of UK producers in the past ten years?*

Figure 4.1 presents two possible time series profiles for the value of a dependent variable (Y).

figure 4.1
Some time series profiles



In Figure 4.1(a) a clear upward trend is evident, around which there is *seasonal variation*, with quarter 1 representing successive peaks and quarter 3 successive troughs within each year. In Figure 4.1(b) a clear downward trend is evident, around which there is a *cyclical component* with a period of around ten years between successive peaks or successive troughs.

Of course many variants are possible for such profiles. We may, for example, have a shorter-term seasonal variation (e.g. quarterly or monthly) superimposed on a time series showing both a cycle and trend using longer-term (e.g. annual) data.

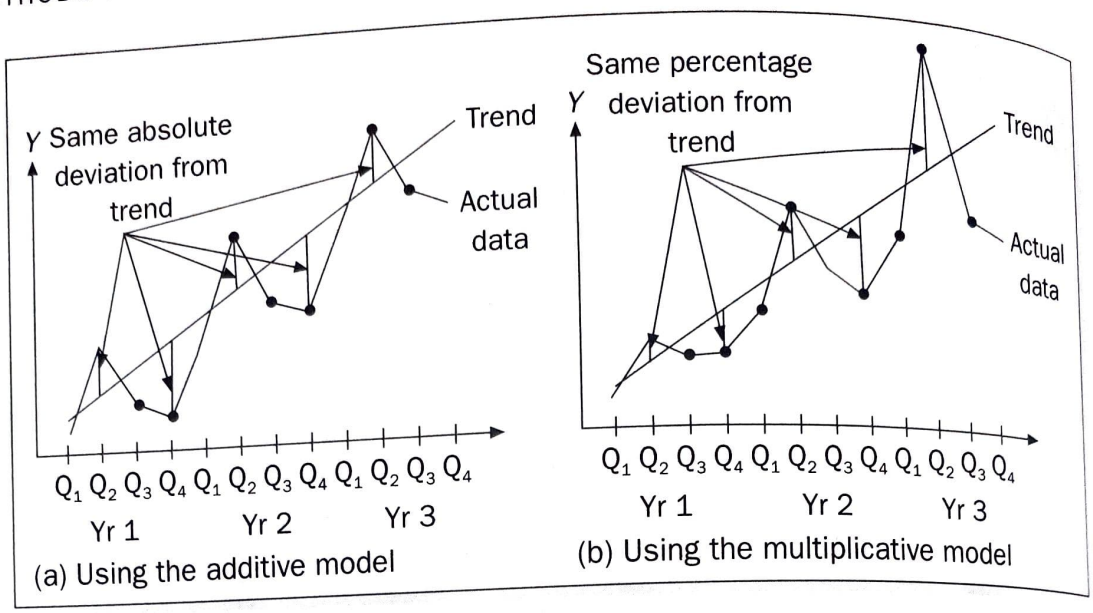
DID YOU KNOW?

The construction industry is widely regarded as a barometer for economic prospects throughout the economy. Traditionally it is one of the earliest sectors to be affected by economic recession and one of the earliest sectors to experience higher demand when economic recovery occurs.

Additive or multiplicative model

We can represent or 'model' our time series using the four components already identified. We can do this in one of two ways; either *adding* or *multiplying* the respective components.

figure 4.2
Choosing an appropriate time series model


Y Same absolute deviation from trend
Trend
Actual data
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
Yr 1
Yr 2
Yr 3
(a) Using the additive model
Y Same percentage deviation from trend
Trend
Actual data
Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4
Yr 1
Yr 2
Yr 3
(b) Using the multiplicative model

- $Y = T + S + C + I$ **Additive model**
- $Y = T \times S \times C \times I$ **Multiplicative model**

As we shall see, we can use either of these models to calculate any trend line or seasonal variation. However there may be circumstances where one approach is more appropriate than another!

As we see in Figure 4.2(a), the **additive model** is the most appropriate where deviations from the trend line are of a similar *absolute magnitude* from one peak (or trough) to another. This is shown in Figure 4.2(a) as occurring for short-term cycles (e.g. seasonal variations, with Q2 providing peaks and Q4 troughs). The same principle would equally apply if we were using annual data and comparing successive peaks or troughs in a business cycle over, say, an eight to ten year period.

However as we see in Figure 4.2(b), the **multiplicative model** is the most appropriate where deviations from the trend line are of a similar *percentage* from one peak (or trough) to another. We can see that the absolute magnitude of the deviations from the trend line grows over time as the dependent variable (Y) rises, but that the percentage deviations remain roughly constant.

4.2 Moving average

As we shall see in the next section, we can use the idea of a **moving average** to find the **trend** of the data. To find the moving average we initially find the simple average (arithmetic mean) for a specified number of items of data. We then recalculate that average having dropped the initial item of data and added a subsequent item of data.

In the worked example below involving quarterly data for sales value, we initially find the simple average for four items of data, and then *move* that average along.

WORKED EXAMPLE **4.1**

Find a four quarter centred moving average for the following data on sales value (£000):

Solution

(1) Year and	Quarter	(2) Sales value (£000)	(3) 4 quarter moving total	(4) 4 quarter moving average	(5) Centred 4 quarter moving average
1998	1	87.5			
	2	73.2			
			314.0	78.5	
	3	64.8			78.9
			316.8	79.2	
	4	88.5			79.6
			319.6	79.9	
1999	1	90.3			80.5
			324.0	81.0	
	2	76.0			81.8
			330.2	82.6	
	3	69.2			83.1
			333.8	83.5	
	4	94.7			83.8
			336.2	84.1	
2000	1	93.9			84.5
			339.0	84.8	
	2	78.4			85.5
			344.6	86.2	
	3	72.0			
	4	100.3			

It is often helpful to find the four quarter centred moving average in stages.

- **Four quarter moving total (column 3).** Here we simply sum the data for the initial four quarters in column (2) to find the four quarter *moving total*. Notice that the moving totals fall in between the actual quarterly data in column (2). For example the first moving total falls in between quarters 2 and 3 for 1998.
- **Four quarter moving average (column 4).** We then divide the respective four quarter *moving totals* in column (3) by 4 to find the four quarter *moving average*.
- **Four quarter centred moving average (column 5).** It will help to *align* the moving averages with the specific quarterly data in column (2). For this reason we *centre* the data in column (4) by summing respective pairs of data and dividing by 2.

DID YOU KNOW?

Survey evidence suggests that the technique of moving averages is used by around 25 per cent of UK businesses in an attempt to forecast future demand.

4.3 Finding the trend

In practice we are often presented with monthly or quarterly data over a number of years for which no obvious business

cycle (C) of eight to ten years is present. We can therefore regard the *actual data*, Y, as having only three of the four components previously mentioned

i.e. $Y = T + S + I$ (Additive)

or $Y = T \times S \times I$ (Multiplicative)

If we calculate an appropriate *moving average* for the monthly or quarterly data, then we can *eliminate* both the seasonal variation component (S) and the irregular component (I) from the actual data. This will leave us with T, the required trend line.

- Using a four quarter moving average in column (4) of the worked example above means that we can *eliminate any seasonal variation* (S) as regards high or low quarters. For example in column (3) after our initial calculation of 314 we drop the first quarter of 1998 but add the first quarter of 1999 to get 316.8. Our four quarter moving total which is the basis for the subsequent centred four quarter moving average in column (5) therefore continually takes in *all* quarters, both high (as for quarter 1) and low (as for quarter 3). In this way column (5) continually eliminates S, the seasonal variation component.
- In calculating columns (4) and (5) we are *averaging*, and any averaging helps eliminate any irregular component. In this way column (5) helps eliminate I, the irregular component.

  If $Y = T + S + I$

  or $Y = T \times S \times I$

then column (5), having eliminated S and I, leaves us with T, the trend line.

Figure 4.3 plots the original quarterly data (Y) of column (2) and the trend line (T) of column (5) for the worked example.

Notice how we use the four quarter centred moving average for the trend line and use straight lines to connect successive points. An alternative approach would, of course, be to draw a single least squares line (see p. 52) which 'best fits' the points which correspond to the trend. We consider this alternative approach below (p. 87).

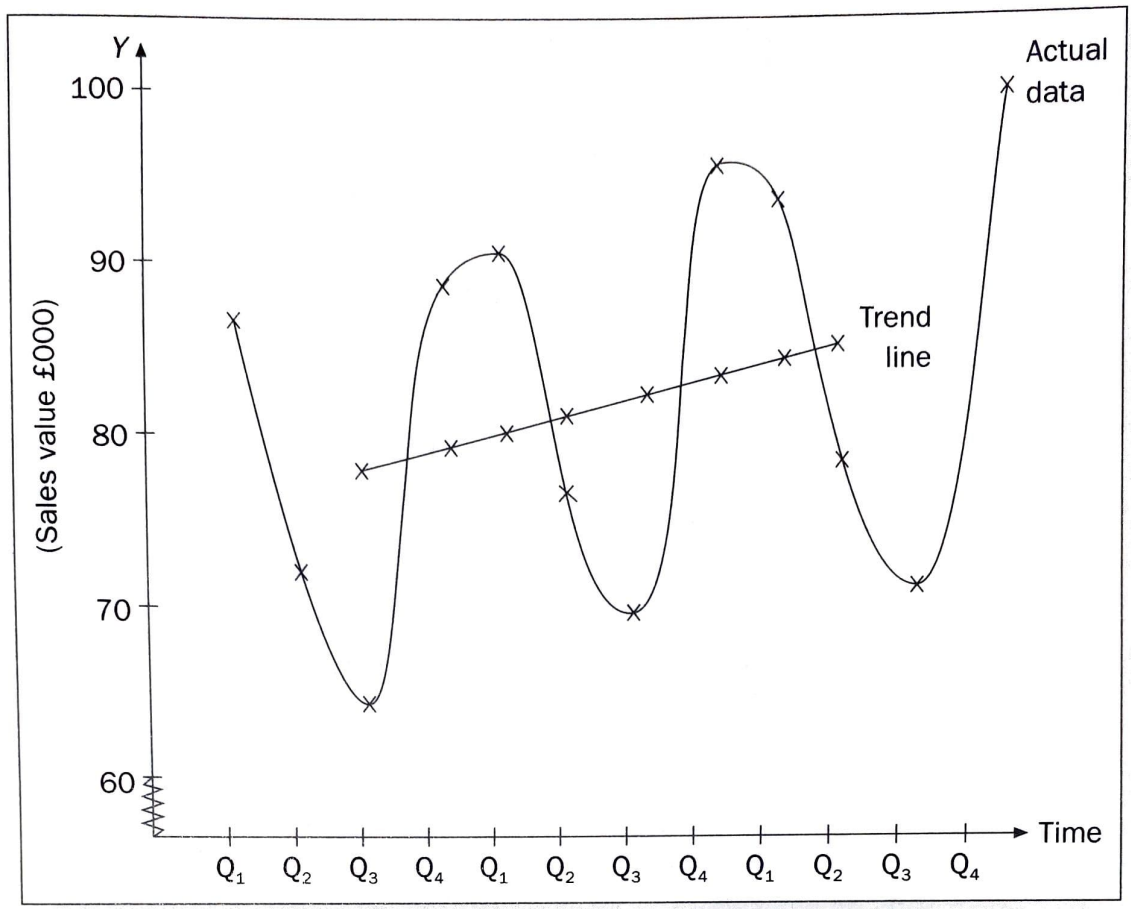
PAUSE FOR THOUGHT 4.3 *Can you name three products (goods or services) for which an upward sales trend is well established, and three products for which a downward sales trend is well established?*

The nature of the data will define the *type* of moving average required to eliminate S and I and therefore leave only the trend line, T.

Using moving averages to find trend (T)

- **Quarterly data**: use four quarter centred moving average to eliminate S and I, and leave T.
- **Monthly data**: use 12 month centred moving average to eliminate S and I, and leave T.

figure 4.3
Plotting the trend line



SELF-CHECK QUESTION

4.1 UK passenger movements by air are shown in the table below (data in 100,000 passengers):

Year	*Q1*	*Q2*	*Q3*	*Q4*
1998	44	80	120	60
1999	52	88	126	62
2000	60	98	140	68

Use an appropriate moving average to find the trend of the data. Plot the original data and your trend line on a graph.

Note: **An answer can be found on p. 334.**

4.4 Finding and eliminating the seasonal variation

When the original data (Y) has no obvious longer-term business cycle in evidence, we have already noted that it can be represented as either:

$$Y = T + S + I \quad \text{(additive model)}$$

or $$Y = T \times S \times I \quad \text{(multiplicative model)}$$

Finding the seasonal variation (S)

We can easily extend our work in using moving averages to find T, the trend component, so that we can find S, the seasonal variation component. Essentially we need only add an extra column to the calculations already undertaken to find T. This extra column which gives us S (and I) will be $(Y - T)$ in the case of the *additive model* or $\left(\frac{Y}{T}\right)$ in the case of the multiplicative model, for the following reasons:

$$Y = T + S + I \text{ (additive model)}$$

$$MA = T$$

$$\underline{Y - MA = S + I}$$

or $Y = T \times S \times I$ (multiplicative model)

$$MA = T$$

$$\underline{\frac{Y}{MA} = S \times I}$$

As we shall see, a further simple averaging process will remove I, leaving S, the seasonal variation which we require.

WORKED EXAMPLE **4.2**

We can demonstrate this approach by using the earlier calculations (p. 81) for sales value in Worked Example 4.1. Here we shall use the *additive* model.

Solution

table 4.1
Time series of quarterly sales value (£000)

		Y (original data)	T Four quarter centred moving average	S + I (Y – T)
1998	Q1	87.5		
	Q2	73.2		
	Q3	64.8	78.9	–14.1
	Q4	88.5	79.6	8.9
1999	Q1	90.3	80.5	9.8
	Q2	76.0	81.8	–5.8
	Q3	69.2	83.1	–13.9
	Q4	94.7	83.8	10.9
2000	Q1	93.9	84.5	9.4
	Q2	78.4	85.5	–7.1
	Q3	72.0		
	Q4	100.3		

Subtracting our trend values (T) from the corresponding quarterly sales values (Y) gives us $S + I$ over successive quarters. The averaging process will then tend to remove I, leaving us with S as required.

A simple table can usefully illustrate this process.

table 4.2
Finding the seasonal variation factors (£000)

	Q1	Q2	Q3	Q4
1998			−14.1	8.9
1999	9.8	−5.8	−13.9	10.9
2000	9.4	−7.1		
Total	19.2	−12.9	−28.0	19.8
average (S)	9.6	−6.45	−14	9.9

The averaging process for each quarter leaves us with S, the seasonal variation for that quarter. Clearly quarters 1 and 4 are high sales quarters for this product, whereas quarter 2 and especially quarter 3 are low sales quarters.

Adjusted S

The data shown in the 'average' row for S are actually *unadjusted*. If we sum the plus values we have +19.5 (9.6 + 9.9) but if we sum the negative values we have −20.45 (−6.45 + −14.0). In other words there is a *net value* for S of −0.95.

Strictly speaking the plus and minus values should cancel out. We can compensate for the −0.95 by adding +0.95/4 (i.e. +0.24) to *each* of the four quarterly values for S. These values would then be the **adjusted values** for S.

table 4.3
Finding the adjusted seasonal variation factors (£000)

	Q1	Q2	Q3	Q4
Unadjusted S (0.95 net)	+9.6	−6.45	−14.0	+9.9
Adjusted $*S$ (zero net)	+9.84	−6.21	−13.76	+10.14

$* + \left(\frac{0.95}{4}\right)$ i.e. +0.24 added to each quarter

It is these adjusted values for S that we usually refer to as the seasonal variation components. Here we can see that sales are normally +9.84 above trend for Q1, but −6.21 below trend for Q2, and so on (all values in £000).

Eliminating the seasonal variations (S)

Having found S, the seasonal variation component, our next step is to *eliminate* S from the original data, Y. We will then have an estimate of Y without any 'distortion' resulting from seasonal influences on Y. Such data with the seasonal variation removed is sometimes called **deseasonalised data**.

table 4.4
Eliminating the seasonal variation from the original data (i.e. deseasonalising the data)

		Y (original data)	*S* (adjusted seasonal variation)	*Y* – *S* (deseasonalised data)
1998	Q1	87.5	+9.84	77.66
	Q2	73.2	–6.21	79.41
	Q3	64.8	–13.76	78.56
	Q4	88.5	+10.14	78.36
1999	Q1	90.3	+ 9.84	80.46
	Q2	76.0	–6.21	82.21
	Q3	69.2	–13.76	82.96
	Q4	94.7	+10.14	84.56
2000	Q1	93.9	+9.84	84.06
	Q2	78.4	–6.21	84.61
	Q3	72.0	–13.76	85.76
	Q4	100.3	+10.14	90.16

Notice that in eliminating S by subtracting S from Y, we sometimes have – – = +, as in Q2 of 1998 when 73.2 – – 6.21 = 79.41.

Using the adjusted seasonal variation component, *S*, calculated above we can easily eliminate *S* from the original data, *Y*. Here we use the *additive* model for illustration purposes and therefore *subtract S* from *Y* for each quarter (in the multiplicative model we would *divide S* into *Y* for each quarter).

ACTIVITY **4**

It can be an involved process calculating the components of time series data. The spreadsheet (p. 467) provides a template for such calculations with the values of the respective quarterly data entered in column D and the calculations and results carried out automatically. Although the layout is for quarterly data, the concept can be adapted for any time series data. The values 2 and two in cells **I1** and **J1** determine the number of decimal places in selected columns.

The layout and formulae for the spreadsheet are shown in Figure A4.2 (p. 467).

As you will see, you are using the technique of a four quarter centred moving average to estimate the trend of the data (had the data been monthly, you would have used a twelve month centred moving average to estimate trend). You are also finding the (adjusted) seasonal factors. If you are using your trend values for *forecasting* you will need to subtract the seasonal factors from each quarterly trend estimate to get a more accurate forecast for that quarter.

Set up your spreadsheet and then enter the quarterly data values shown below.

Year	Q1	Q2	Q3	Q4
Year 1	196.9	295.5	349.4	389.3
Year 2	324.1	418.0	447.5	456.4
Year 3	415.2	528.6	550.6	615.3
Year 4	513.6	650.8	670.6	754.4

- Comment on your results
- Check your solution with the response on p. 466
- Repeat the procedure using the quarterly data values in the various Self-check and Review questions involving moving averages.

4.5 Forecasting: time series

We have already looked at prediction (forecasting) using the simple regression analysis of Chapter 3 (p. 52). Here we *combine* the idea of finding a least squares line (line of 'best fit') with our analysis of time series components. We can then *extrapolate* (take forward) the regression line which best fits the trend estimates to derive a forecast of future sales, etc.

Forecasting: trend value only

For example we can fit a single *least squares line* to the trend estimates on the earlier scatter diagrams (e.g. Figure 4.3, p. 83), instead of merely linking successive trend estimates by a sequence of straight lines. This regression line, once calculated, can then be used for prediction (forecasting) of future trend values.

We must initially calculate the regression line for the eight trend estimates shown in Figure 4.3 and Table 4.1. We can use *either* of the two approaches considered in Chapter 3 for calculating a regression line. Here we use the 'original data' formula (p. 53), with the dependent variable Y standing for the trend value and the independent variable X standing for time (quarters). With trend estimates available for eight successive quarters, we let $X = 1$ for the first available quarter for a trend estimate (Q2, 1998), rising to $X = 8$ for the eighth available quarter for a trend estimate (Q3, 2000).

From Table 4.1 above (p. 84) we can then construct a new table allowing the application of our regression formula:

X *(quarter of data)*	Y *(trend value)*	XY	X^2	Y^2
1	78.9	78.9	1	6,225.2
2	79.6	159.2	4	6,336.2
3	80.5	241.5	9	6,480.3
4	81.8	327.2	16	6,691.2
5	83.1	415.5	25	6,905.6
6	83.8	502.8	36	7,022.4
7	84.5	591.5	49	7,140.3
8	85.5	684.0	64	7,310.3
$\Sigma X = 36$	$\Sigma Y = 657.7$	$\Sigma XY = 3{,}000.6$	$\Sigma X^2 = 204$	$\Sigma Y^2 = 54{,}111.5$

$$\bar{X} = \frac{36}{8} = 4.5 \quad \bar{Y} = \frac{657.7}{8} = 82.2$$

We noted (p. 53) that, using original data, the formula for the regression (least squares) line is:

$$\hat{Y} = mX + c$$

where $m = \dfrac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$

and $c = \bar{Y} - m\bar{X}$

where n = the number of observations

i.e. $m = \dfrac{8(3{,}000.6) - (36)(657.7)}{8(204) - (36)^2}$

$$m = \frac{24{,}004.8 - 23{,}677.2}{1{,}632 - 1{,}296}$$

$$m = \frac{327.6}{336}$$

$$\underline{m = +0.975}$$

$$c = \bar{Y} - m\bar{X}$$

i.e. $c = 82.2 - 0.975(4.5)$

$$\underline{c = 77.8}$$

$$\underline{\hat{Y}_T = 0.975X + 77.8}$$

where $\hat{Y}_T$ = trend sales value (£000)

X = quarterly time period (Q2, 1998 = 0)

We can now use this regression line to **predict** or **forecast** future trend values should this past relationship hold into the future.

Figure 4.4 plots this regression line for the trend and extrapolates the trend line into the future.

The dashed part of the trend line indicates where the extrapolation occurs, i.e. where the line is extended beyond that for which data are currently available.

We can use the regression line for forecasting. For example we can see from Figure 4.4 that:

when X = 12 (i.e. Q2, 2001)

$$\hat{Y}_T = 0.975(12) + 77.8$$

$$\hat{Y}_T = 11.7 + 77.8$$

i.e. $\underline{\hat{Y}_T = 89.5 \text{ (to 1 d.p.)}}$

Similarly, when X = 20 (i.e. Q2, 2003)

$$\hat{Y}_T = 0.975(20) + 77.8$$

$$\hat{Y}_T = 19.5 + 77.8$$

$$\underline{\hat{Y}_T = 97.3 \text{ (to 1 d.p.)}}$$

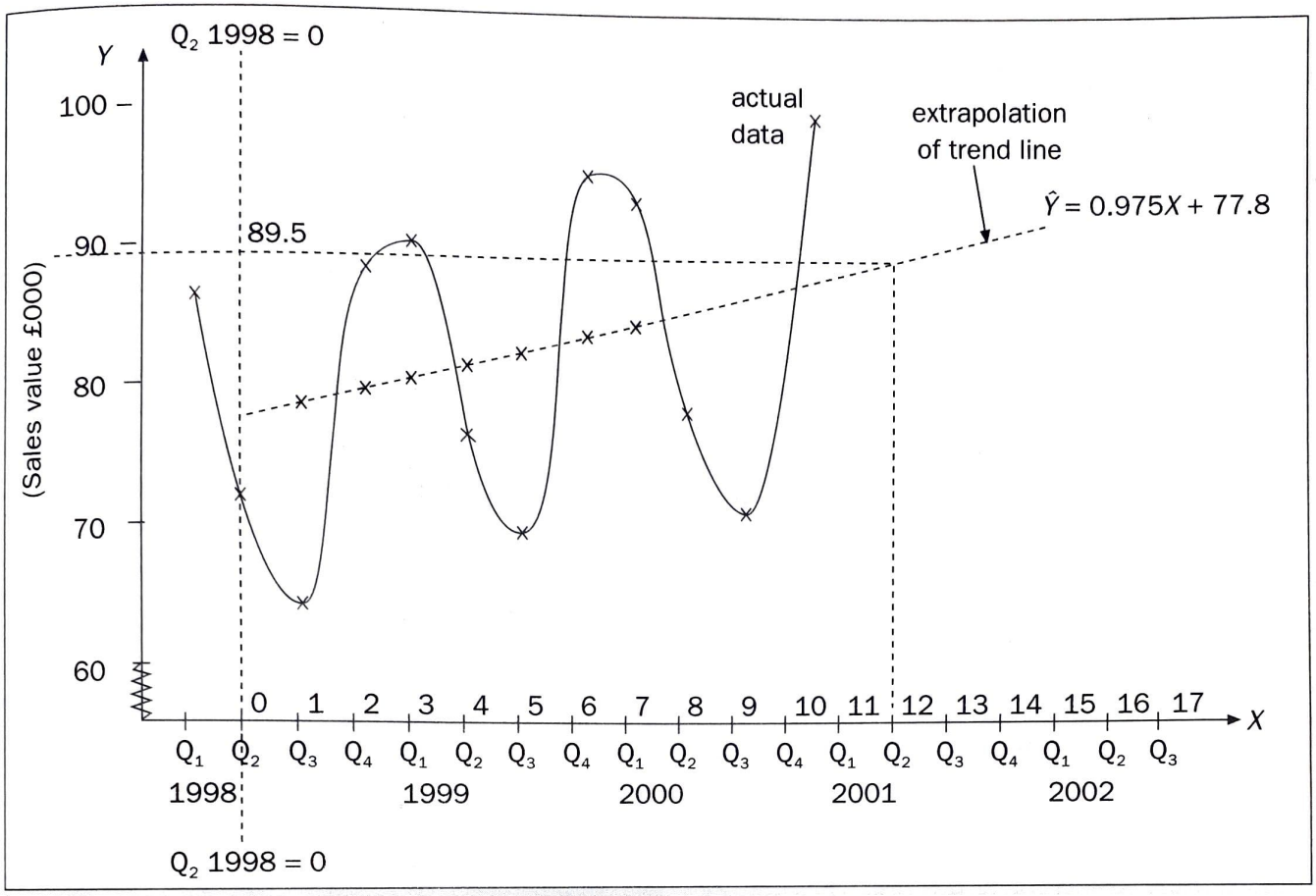


figure 4.4
Using the regression analysis to find the trend line and using that line for forecasting

Of course, the better the fit of our regression line to the existing trend data, the more confident we can be in extrapolating the line to make forecasts in the future.

This brings us back to the coefficients of determination and correlation (Chapter 3, pp. 58–68). In this particular case there is an extremely good fit for our trend regression line.

Using original data

$$R^2 = \left[\frac{n\sum XY - \sum X\sum Y}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right] \times \left[n\sum Y^2 - \left(\sum Y\right)^2\right]}}\right]^2$$

$$\text{i.e. } R^2 = \left[\frac{8(3{,}000.6) - (36)(657.7)}{\sqrt{\left[8(204) - (36)^2\right] \times \left[8(54{,}111.5) - (657.7)^2\right]}}\right]^2$$

$$\text{i.e. } R^2 = \left[\frac{24{,}004.8 - 23{,}677.2}{\sqrt{[1{,}632 - 1{,}296] \times [432{,}892.0 - 432{,}569.3]}}\right]^2$$

$$R^2 = \left[\frac{327.6}{\sqrt{336 \times 322.7}}\right]^2 = \left[\frac{327.6}{329.3}\right]^2 = [0.995]$$

R^2 = coefficient of determination = 0.995

R = coefficient of correlation = 0.997

With over 99 per cent of the total variation 'explained' or 'accounted for' by our regression line (see p. 60), we can have considerable confidence in forecasting into the future by extrapolating from our regression line.

Forecasting: trend and seasonal factors

Of course any forecast for the sales in a particular quarter will be more accurate if it takes into account the *seasonal variation*, *S*, components already identified as well as the trend. The adjusted values for *S* were as follows (£000):

Q1 (+9.84) Q2 (–6.21) Q3 (–13.76) Q4 (+10.14)

We must add or subtract these seasonal variation values to our trend forecast if we are to be more realistic about the likely future sales in a particular quarter.

For example, we can forecast the trend using the equation

$$\hat{Y}_T = 0.975X + 77.8$$

where $X = 0$ for Q2, 1998.

This gives the *trend* forecast for Q2, 2001 as follows:

▶ when $X = 12$ (i.e. Q2, 2001)

$$\hat{Y}_T = 0.975\ (12) + 77.8$$

$$\underline{\hat{Y}_T = 89.5}$$

However Q2 also has a *seasonal* component

$$\underline{S = -6.21}$$

So our forecast for $\hat{Y}_{T+S}$ with *both* T and S included is:

$$\hat{Y}_{T+S} = 89.5 - 6.21$$

$$\underline{\hat{Y}_{T+S} = 83.29\ (£000)}$$

Table 4.5 shows the forecast ($\hat{Y}$) for sales value (£000) over the period 2001-2003 inclusive.

table 4.5

Forecasts of $\hat{Y}_{T+S}$ including trend (T) and seasonal variation (S) components. Additive model with least squares line $\hat{Y}_T = 0.975X + 77.8$ for trend forecast (Q2, 1998 = 0).

Year and quarter		*X**	*($\hat{Y}_T$) Trend forecast*	*(S) Seasonal variation*	*$\hat{Y}_{T+S}$ Total forecast*
2001	Q1	11	88.5	+9.84	98.34
	Q2	12	89.5	–6.21	83.29
	Q3	13	90.5	–13.76	76.74
	Q4	14	91.5	+10.14	101.64
2002	Q1	15	92.4	+9.84	102.24
	Q2	16	93.4	–6.21	87.19
	Q3	17	94.4	–13.76	80.64
	Q4	18	95.4	+10.14	105.54
2003	Q1	19	96.4	+9.84	106.24
	Q2	20	97.3	–6.21	91.09
	Q3	21	98.3	–13.76	84.54
	Q4	22	99.3	+10.14	109.44
	etc.				

Note: * Quarter 2 1998

SELF-CHECK QUESTION

4.2 The data below show UK sales of a particular model of car over three years (figures in 100,000 units).

Quarter	*1*	*2*	*3*	*4*
1998	66	106	140	82
1999	73	119	165	91
2000	85	130	205	100

(a) Use the technique of moving averages to find values for the trend component (*T*).
(b) Plot the original data and your trend values on a scatter diagram.
(c) Find a 'least squares line' to best fit your trend values.
(d) Estimate the season variation component (*S*).
(e) Forecast future sales, taking both trend and seasonal variation into account in:
(i) Quarter 1, year 2003
(ii) Quarter 3, year 2003
(f) How confident can you be in these forecasts?

***Note:* An answer can be found on p. 335.**

4.6 Forecasting: exponential smoothing

A widely used technique in business forecasting is that of **exponential smoothing**. The idea here is that the forecast at time t for the next time period $(t + 1)$ should take into account the *observed error* in the forecast made for t in the previous time period $(t - 1)$.

Clearly such 'exponential smoothing' implies a learning process, whereby future forecasts are continually revised (smoothed) in the light of previous experience. Strictly speaking, this approach is most appropriate when there is little or no trend (T) in the data and little or no seasonal variation (S). It is best used for short-term forecasting, for example forecasting the outcome of the next time period.

We can express this approach as follows:

$$\text{Next forecast} = \text{Previous forecast} + \text{Some proportion of the previous forecasting error}$$

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t)$$

where Y_t = actual data (observed outcome) in time period t.
$\hat{Y}_t, \hat{Y}_{t+1}$ = forecast data for next time period (t or $t + 1$) made in previous time period (t – 1 or t respectively)
α = smoothing constant

Put another way

$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(E_t)$

where E_t = error term = $Y_t - \hat{Y}_t$

The value assigned to α, the **smoothing constant**, can vary between zero and one

i.e. $0 \leq \alpha \leq 1$

We can usefully consider the extreme values of α by way of illustration.

- When $\alpha = 0$, no adjustment is made for the previous forecasting error. The next forecast is then assumed to be the same as the previous forecast.
- When $\alpha = 1$, full adjustment is made for the previous forecasting error. The next forecast is then the previous forecast ± the entire amount of any previous forecasting error.

In practice α can be derived *experimentally*, based on the average size of the error term in previous forecasts. Values of between 0.1 and 0.3 are typically assigned to α.

In the Table 4.6 below we continue this exponential forecast into other quarters, based on the actual values of Y (quarterly sales) shown (000 units)

table 4.6
Exponential forecast of future demand (000 units), $\alpha = 0.2$

Quarter	Y_t (actual sales)	$\hat{Y}_t$ (forecast sales)	E_t $(Y_t - \hat{Y}_t)$
1	96		
2	102	84	18
3	104	87.6	16.4
4	100	90.9	9.1
5	88	92.7	−4.7
6	92	91.8	0.2
7	96	91.8	4.2
8	94	92.6	1.4
9	90	92.9	−2.9
10	96	92.3	3.7

WORKED EXAMPLE **4.3**

A firm is at the end of quarter 2. At the end of quarter 1 it had forecast that demand for its product in quarter 2 would be 84,000 units, only to find that actual demand turned out to be higher at 102,000 units. Provide an exponential forecast for demand in quarter 3, using $\alpha = 0.2$.

Solution

Y_2 = 102,000 (actual demand Q2)

$\hat{Y}_2$ = 84,000 (forecast demand Q2)

Here t, the current time period, is quarter 2 and α is 0.2.

$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t)$

i.e. $\hat{Y}_{2+1} = \hat{Y}_2 + 0.2(Y_2 - \hat{Y}_2)$

$\hat{Y}_3 = 84{,}000 + 0.2\ (102{,}000 - 84{,}000)$

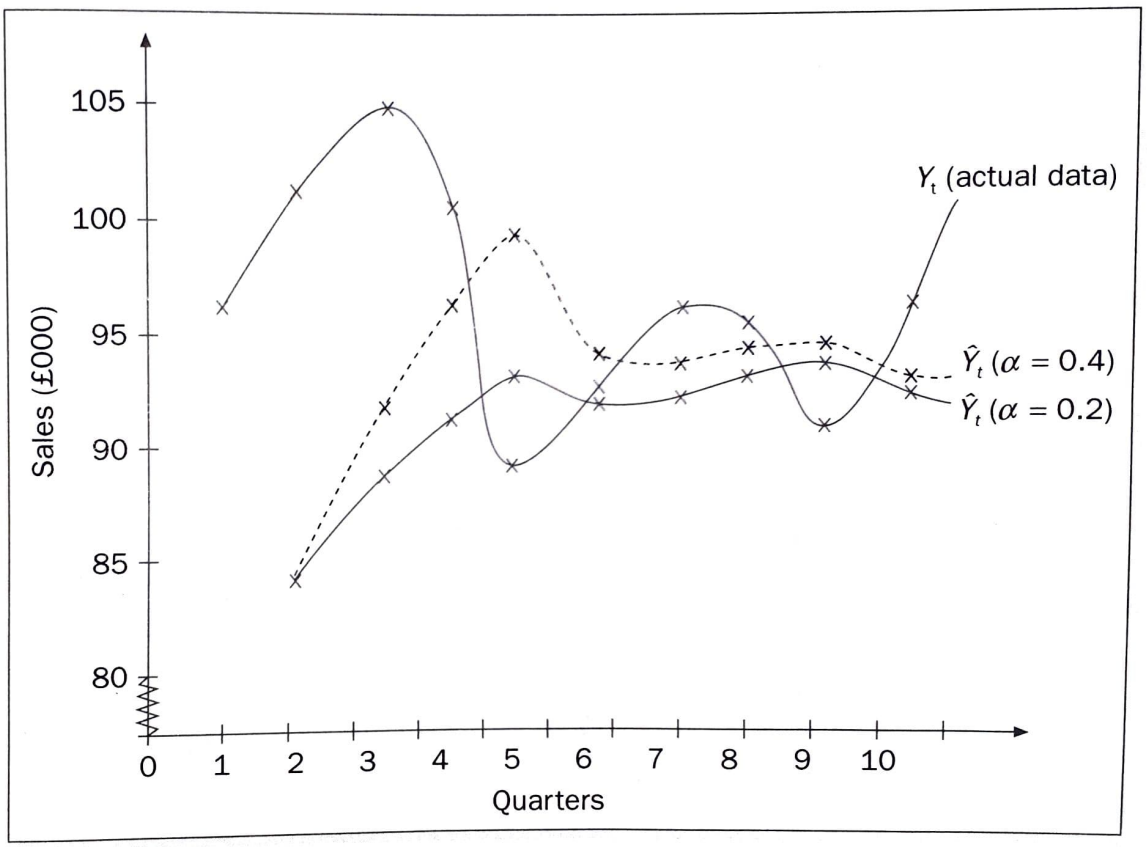
$\hat{Y}_3 = 84{,}000 + 0.2\ (18{,}000)$

$\hat{Y}_3 = 84{,}000 + 3{,}600$

$\underline{\hat{Y}_3 = 87{,}600}$

The impact of this exponential smoothing with $\alpha = 0.2$ on forecast sales is shown in Figure 4.5.

figure 4.5
Forecasting future sales using exponential smoothing and different values for α



Different values of α

In Table 4.6 we assumed the 'smoothing constant', α, to be 0.2. We have already looked at the impact of extreme values ($\alpha = 0$, $\alpha = 1$) of α on exponential smoothing. In general we can say the following.

- When α is close to 0, only a small proportion of the latest error (E_t) will be included in the next forecast ($\hat{Y}_{t+1}$). Clearly that next forecast will not differ greatly from the previous forecast.
- When α is close to 1, a large proportion of the latest error (E_t) will be included in the next forecast ($\hat{Y}_{t+1}$). Clearly that next forecast will differ greatly from the previous forecast.

As we can see from Figure 4.5, using $\alpha = 0.2$ gives a smoother sequence of forecasts than would occur using $\alpha = 0.4$. This is because the $\alpha = 0.2$ sequence

of forecasts for $\hat{Y}_{t+1}$ is less influenced by any actual observation (Y_t) and its divergence from its forecast value $(\hat{Y}_t)$.

Choosing a value for α

It would seem reasonable to choose a value of α which, on previous experience, would have produced the 'best' forecast using exponential smoothing. 'Best' in this context would arguably be the value of α which gave a smaller sequence of errors (E_t), where $E_t = Y_t - \hat{Y}_t$.

Two possible ways of evaluating previous forecasts suggest themselves.

Mean absolute error (MAE)

Here we select that value of α which, on previous experience, minimises the average of the *absolute errors* (ignoring sign) of deviations between actual (Y_t) and forecast $(\hat{Y}_t)$ values:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |E_t|}{n}$$

where | | is modulus, i.e. ignore signs
n is number of error terms $(Y_t - \hat{Y}_t)$

Notice that n, the number of error terms, will be one less than the number of items of data since we do not have a forecast and therefore error term for the first item of data.

Mean squared error (MSE)

Here we select that value of α which, on previous experience, minimises the average of the *squares* of the errors. We have no sign problem here, as all squared values are positive:

$$\text{MSE} = \frac{\sum_{i=1}^{n} E_t^2}{n}$$

where n is number of error terms $(Y_t - \hat{Y}_t)$

WORKED EXAMPLE **4.4**

Use the data of Table 4.6 to provide forecasts using $\alpha = 0.2$, 0.3 and 0.4 respectively. Which value of α would you select for future forecasts?

Solution

We have already plotted the data for $\alpha = 0.2$ and 0.4 on Figure 4.5. Here we formally work out MAE and MSE for the nine error terms (E_t) based on these ten items of data.

		$\alpha = 0.2$		$\alpha = 0.3$		$\alpha = 0.4$	
Quarter	Y_t	E_t	E_t^2	E_t	E_t^2	E_t	E_t^2
1	96						
2	102	18.0	324.0	18.0	324.0	18.0	324.0
3	104	16.4	269.0	14.6	213.2	12.8	163.8
4	100	9.1	82.8	6.2	38.4	3.7	13.7
5	88	−4.7	22.1	−7.6	57.8	−9.8	96.0
6	92	0.2	0.0	−1.4	2.0	−1.9	3.6
7	96	4.2	17.6	3.1	9.6	2.9	8.4
8	94	1.4	2.0	0.1	0.0	−0.3	0.1
9	90	−2.9	8.4	−3.9	15.2	−4.2	17.6
10	96	3.7	13.7	3.3	10.9	3.5	12.3
		$\Sigma\|E_t\| = 60.6$	$\Sigma E_t^2 = 739.6$	$\Sigma\|E_t\| = 58.2$	$\Sigma E_t^2 = 671.1$	$\Sigma\|E_t\| = 57.1$	$\Sigma E_t^2 = 639.5$

Note: Although there are ten items of data there are only nine error terms, hence $n = 9$.

$$\text{MAE} = \frac{\sum |E|_t}{n}$$

(a) $\alpha = 0.2 = \dfrac{60.6}{9} = 6.73$

(b) $\alpha = 0.3 = \dfrac{58.2}{9} = 6.47$

(c) $\alpha = 0.4 = \dfrac{57.1}{9} = 6.34$

$$\text{MSE} = \frac{\sum E_t^2}{n}$$

(a) $\alpha = 0.2 = \dfrac{739.6}{9} = 82.18$

(b) $\alpha = 0.3 = \dfrac{671.1}{9} = 74.57$

(c) $\alpha = 0.4 = \dfrac{639.5}{9} = 71.06$

In this particular example the value $\alpha = 0.4$ will give the minimum value for *both* mean absolute error (MAE) and mean square error (MSE). It follows that our previous experiences of forecasting would arguably lead to the choice of $\alpha = 0.4$ for future forecasts when choosing between the possible values of $\alpha = 0.2$, 0.3 or 0.4.

REVIEW QUESTIONS

4.1 As a result of local government cutbacks some branch libraries are threatened with closure. A newly appointed chief librarian claims that a certain branch library has a declining patronage, based on his observations of the last three months. The head librarian of the branch disagrees with the analysis and quotes quarterly data recorded over the past three years for book issues to support his case. That data are as follows:

Year	Q1	Q2	Q3	Q4
1998	2,542	2,826	2,991	2,644
1999	2,766	2,905	3,048	3,137
2000	2,944	3,140	3,333	3,125

(a) Calculate the trend line using the moving average method.
(b) Plot the original data and the trend values on a scatter diagram.
(c) Calculate the seasonal variation factors.

4.2 The manufacturer of snow chains is considering building a new production plant to cope with growing demand. A tyre firm is interested in investing in the company and has been provided with the following quarterly sales figures:

Unit sales 1997–2000 (000s)

Year	Q1	Q2	Q3	Q4
1	441.1	397.7	396.1	472.8
2	476.4	454.4	450.8	553.5
3	580.7	573.2	571.6	703.6
4	692.0	676.5	659.9	752.8

(a) Calculate the trend line using the moving average method.
(b) Plot the original data and the trend values on a scatter diagram.
(c) Estimate and then eliminate the season variation component. In other words find the deseasonalised data.

4.3 A company producing torches is negotiating with a company for the supply of torch bulbs. The bulb company therefore needs to plan its production to meet the needs of the torch company and thus uses that company's quarterly sales figures over the past three years to forecast future demand. The sales figures are as follows:

Quarterly sales figures 1998–2000 (000s)

Year	Q1	Q2	Q3	Q4
1	349.4	295.5	196.9	389.3
2	447.5	418	324.1	456.4
3	550.6	528.6	415.2	615.3

Using the moving average technique find the trend observations and the seasonal variation factors. Fit a least squares regression line to the trend observations. Use this trend regression line and your estimates of the seasonal variation factors to forecast future demand for the four quarters of year 4 for torches.

4.4 The area sales manager of a company is responsible for providing a forecast for the value of sales. However, she has been taken ill and the deputy has now to make the forecast for the next month. The exponential smoothing method, favoured by the company, uses an α of 0.2. However, the deputy believes that an α of 0.3 is a more realistic smoothing constant. Using the data provided, forecast for the next month using each smoothing constant.

t	Y
1	3.85
2	3.22
3	3.18
4	3.61
5	3.62
6	3.48
7	3.53
8	3.62
9	3.36
10	3.58
11	3.62
12	3.24

Calculate the mean absolute error and the mean square error for each smoothing constant (assume an initial forecast for t=1 of 3.49). Which smoothing constant should be recommended based on the criterion of minimising the mean square error?

Note: Answers can be found on pp. 402–415.

Further study and data

Texts

Bancroft, G. and O'Sullivan, G. (1993), *Quantitative methods for accounting and business studies*, 3rd edn, McGraw Hill, chapter 10.

Curwin, J. and Slater, R. (1996), *Quantitative methods for business decisions*, 4th edn, International Thompson Business Press, chapter 20.

Lawson, M., Hubbard, S. and Pugh, P. (1995), *Maths and statistics for business*, Addison Wesley Longman, chapter 11.

Morris, C. (1999), *Quantitative approaches in business studies*, 5th edn, Pitman, chapter 16.

Oakshott, L. (1996), *Essential elements of quantitative methods*, BPP, chapter 3.

Swift, L. (2001), *Quantitative methods for business, management and finance*, Palgrave Publishers, part S.

Thomas, R. (1997), *Quantitative methods for business studies*, Prentice Hall, chapter 6.

Waters, D. (1997), *Quantitative methods for business*, 2nd edn, Addison Wesley Longman, chapter 9.

Wisniewski, M. with Stead, R. (1996), *Foundation quantitative methods for business*, Pitman, chapter 13.

Sources of information and data

See the list at the end of Chapter 1 (p. 25).

On-line exercises

Check the web site *www.booksites.net/burton* to find extra questions, Spreadsheet and on-line exercises for this chapter.

CHAPTER 5

Probability

Objectives

When you have read this chapter you should be able to:

- distinguish between *experimental* and *theoretical* probability;
- apply the 'AND' rule to *independent events* but the 'OR' rule to *mutually exclusive* events;
- use *Venn diagrams* to solve problems where events are not mutually exclusive;
- solve problems involving *conditional probability*, including those where *Bayes theorem* can be used;
- use *decision tree* analysis to show all the probabilities in complex situations, often involving conditional probabilities;
- identify situations in which probability calculations can aid the decision-making process, as in the case of *expected values*;
- calculate probabilities involving *permutations* or *combinations* of events.

Introduction

In this chapter we look at the basis for calculating the **probabilities** with which particular events, or combinations of events, might be expected to occur. In fact probabilities, together with the probability distributions we consider in Chapters 6 and 7, are the basis for many economic and business decisions. For example *insurance* against all kinds of possible future events is only possible because actuaries have worked out, often using past data, the probabilities of those events occurring. These probabilities then determine the *insurance premiums* to be charged for cover against such events. Of course if circumstances change, and with them the associated probabilities of such events occurring, then insurance premiums will have to rise or fall. The greater likelihood of burglary has, for instance, caused a rise in premiums for insuring house contents against theft.

In this chapter we consider the basis for probability calculations involving both **mutually exclusive** ('OR' rule) and **independent** ('AND' rule) events. In fact some events are neither mutually exclusive nor independent, in which case the **Venn** diagram can help us solve probability problems. We then move on to consider **conditional probabilities**, as in the case of Bayes' theorem. Decision tree analysis is seen as being particularly helpful in showing *all* the probabilities. The idea of **expected value** is also explored and applied to practical decision-making situations, using decision tree analysis in some cases. The chapter concludes by looking at **permutations** and **combinations** of events, which is the background to work on the Binomial and Poisson distributions in Chapter 6.

Answers to the 'Self-check questions', responses to each 'Pause for thought' and answers to the 'Review questions' can be found at the end of the book.

5.1 Probability calculations

Experimental probability

This can be found by performing an *experiment* many times and keeping an accurate record of the results. The *experimental probability* of a particular event happening can then be worked out as follows:

$$\text{Experimental probability} = \frac{\text{number of times the event has happened}}{\text{total number of possible occurrences}}$$

Example

When a normal dice was rolled 100 times, the number five was actually rolled 19 times. This gives an *experimental probability* ratio of $\frac{19}{100} = 0.19$.

DID YOU KNOW?

Experimental probabilities are worked out for all types of 'hazards'. For example the 'risk' of dying from occupational exposure to asbestos has been worked out as 3 per 1,000 persons exposed. Actuaries work out experimental probabilities all the time in order to calculate insurance premiums for road accidents by age and other personal characteristics.

Theoretical probability

This is found by considering *equally likely* events. The theoretical probability is found by using the following ratio.

$$\text{Theoretical probability} = \frac{\text{number of ways an event can occur}}{\text{total number of different equally likely events that can occur}}$$

Examples

1 Rolling a dice and getting an even number has theoretical probability $\frac{3}{6} = \frac{1}{2}$

2 Rolling a dice and getting a number 5 has theoretical probability $\frac{1}{6} = 0.17$ (2 d.p)

Of course if an event is *impossible*, then its theoretical probability is 0. On the other hand, if the event is *certain*, then its theoretical probability is 1.

Notice that if we know the theoretical probability of an event occurring, then we also know the theoretical probability of it *not* occurring.

If the probability of an event occurring is, say, p then the probability of the event *not* occurring is $1 - p$.

Example

The theoretical probability of rolling a 5 on a dice $= \frac{1}{6}$

The theoretical probability of *not* rolling a 5 on a dice $= 1 - \frac{1}{6} = \frac{5}{6}$

PAUSE FOR THOUGHT 5.1 *Can you work out the theoretical probabilities for three different events?*

SELF-CHECK QUESTIONS

5.1 Write down the theoretical probability of the following events:

(a) rolling a dice and getting a 2
(b) rolling a dice and not getting a 3
(c) rolling a dice and getting an odd number
(d) rolling a dice and getting a 2 or a 3
(e) drawing an ace from a pack of cards
(f) drawing a card from a pack of cards and not getting a diamond
(g) being born on 31 April
(h) rolling a dice and getting a number less than 7
(i) buying a box of matches with an even number of matches in it (no quality control)
(j) picking a blue ball from a bag containing two blue and seven white balls
(k) picking a ball that is not blue from the above bag
(l) meeting someone whose birthday is not on a Saturday.

5.2 A bag contains 20 marbles, all the same size. Ten are blue, five are white, three are yellow and two are black. What is the theoretical probability of choosing at random

(a) a blue marble
(b) a black marble
(c) a white marble
(d) a marble that is yellow or black
(e) a marble that is not white

5.3 Two unbiased dice are thrown. Their scores are added to make a total between 2 and 12.

(a) There are 36 possible totals as indicated in Figure 5.1 Copy and complete the table.

figure 5.1

one dice \ other dice	1	2	3	4	5	6
6	(1,6)					
5	(1,5)					
4	(1,4)					
3	(1,3)					
2	(1,2)	(2,2)	(3,2)			
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)

(b) Write down the theoretical probability of throwing a total of

(i) 2 (ii) 3 (iii) 4
(iv) 6 (v) 7 (vi) 8
(vii) 10 (viii) 12

(c) What is the theoretical probability of not getting a total of 7?

Note: **Answers can be found on p. 338.**

5.2 Mutually exclusive events: OR rule

If one event excludes the possibility of another event happening, then we say that the two events are **mutually exclusive**

Example (i) tossing a coin and getting a head and a tail;
(ii) rolling a dice once and getting a 2 and a 3.

Events that are mutually exclusive can have their probabilities *added* to find the probability of one **OR** the other happening.

> **OR rule: Mutually Exclusive Events**
> For two mutually exclusive events A and B, the probability of A OR B occurring can be written as:
>
> P(A) OR P(B) = P(A) + P(B)

Example

If I roll a dice what is the probability that I roll a three or four?

$$\text{Probability of rolling a three} = \frac{1}{6}$$

$$\text{Probability of rolling a four} = \frac{1}{6}$$

$$\text{The probability of rolling a three OR a four} = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

WORKED EXAMPLE **5.1**

A stockbroker uses past experience to work out the probabilities of a particular share changing in price over the year. His findings for 'events' A to E and associated experimental probabilities, are as follows:

A Share price rises by 15% or more: P(A) = 0.2
B Share price rises *less than* 15%: P(B) = 0.3
C Share price remains unchanged: P(C) = 0.1
D Share price falls *less than* 15%: P(D) = 0.25
E Share price falls by 15% or more: P(E) = 0.15

Find:

(i) the probability that the share price will rise
(ii) the probability that the share price will not rise
(iii) the probability that the share price will fall

Solution

The events A to E are mutually exclusive, if one happens the others do not:

(i) P(A or B) = P(A) + P(B) = 0.2 + 0.3 = 0.5
(ii) 1 – P(A or B) = 1 – 0.5 = 0.5
(iii) P(D or E) = P(D) + P(E) = 0.25 + 0.15 = 0.40

PAUSE FOR THOUGHT 5.2 *Can you identify three different situations involving mutually exclusive events and calculate the probability using the OR rule?*

5.3 Venn diagrams: events not mutually exclusive

Sometimes events can occur at the same time, so that they are *not* mutually exclusive. In solving this type of problem it can help to make use of a **Venn diagram**.

Suppose, for example, that we have a pack of 52 cards and that event A is drawing a club and event B drawing a picture card. We now want to work out the probability of a card drawn at random being either a club or a picture card, i.e. P(A OR B). A pack of 52 cards will have 13 clubs and 12 picture cards.

In this case we *cannot* use the OR rule for mutually exclusive events since here a club could itself be a picture card, i.e. events A and B are *not* mutually exclusive. Of the 13 club cards in the pack, the Jack, Queen and King of clubs are picture cards. There is therefore a danger of double counting if we use the conventional OR rule for mutually exclusive events. The correct answer here is:

$$P\ (A\ OR\ B) = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52}$$

OR rule: Non-Mutually Exclusive Events

P(A OR B) = P(A) + P(B) – P(AB)

The Venn diagram (Figure 5.2) can be useful in solving problems involving events that are *not* mutually exclusive, i.e. which 'overlap' in some way.

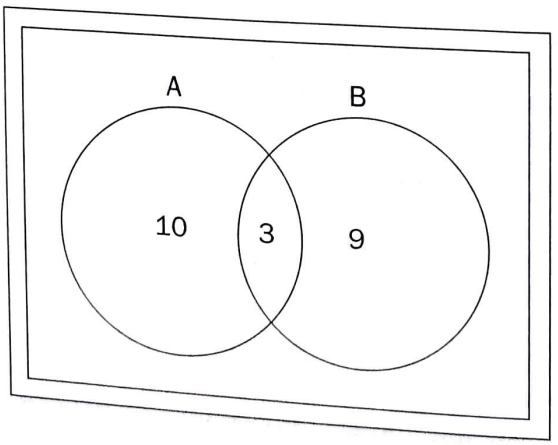


figure 5.2
Venn diagram for pack of cards, where A = club card and B = picture card

We can note the following from Figure 5.2.

- The area of the rectangle corresponds to *all* the possible events in the sample, here the 52 different cards which could be drawn from the pack.
- The circle A corresponds to the 13 club cards.
- The circle B corresponds to the 12 picture cards.
- The *overlap* between the two circles corresponds to the fact that three of the club cards can themselves be picture cards.

When completing a Venn diagram it is essential that you first consider the *area of overlap* before going any further. Therefore enter the 3 in the area of overlap in Figure 5.2, then 10 in the remainder of the circle A representing club cards and 9 in the remainder of the circle B representing picture cards. The required probability that a card drawn at random from the pack is a club OR a picture is obtained by adding up these numbers, namely 10 + 3 + 9 = 22, and then dividing by the total sample of 52 to obtain the result (22/52) already outlined.

SELF-CHECK QUESTIONS

5.4 A company puts in two separate bids, A and B, for a particular contract. The probability that it will obtain the contract by bid A is 0.3, but 0.5 that it will obtain the contract by bid B. What is the probability that it will obtain the contract by bid A or bid B?

5.5 A company usually recruits 25 management trainees per annum. In 2001, 140 applications were received and of these:

78 had previous work experience
43 had passed a vocational exam
21 had both work experience and had passed a vocational exam (and had been included in the above figures)

(a) Use a Venn diagram to illustrate this situation.
(b) What is the probability that an applicant selected at random had previous work experience or had passed a vocational exam?

Note: **Answers can be found on pp. 338–339.**

5.4 Independent events: AND rule

If *one* event has no effect on another event then we say that the events are **independent**. To find the probability of BOTH events happening (one event **AND** then the other) we *multiply* their probabilities.

Example

From a normal pack of cards, what is the probability of drawing an Ace, putting it back and then drawing a King?

Probability of drawing an Ace $\frac{4}{52} = \frac{1}{13}$

Probability of drawing a King $\frac{4}{52} = \frac{1}{13}$

Since we have independent events, we multiply their individual probabilities,

i.e. $\frac{1}{13} \times \frac{1}{13} = \frac{1}{169}$

Events that are independent can have their probabilities *multiplied* to find the probability of one AND the other happening.

AND rule
For two independent events A and B, the probability of A AND B occurring can be written as:

P(A) AND P(B) = P(A) × P(B)

WORKED EXAMPLE **5.2**

Each one of three separate boxes has 100 electrical components in it. The first box has six faulty items, the second 8 and the third 5. One component is taken from each box. What is the probability that:

(a) all three components are faulty
(b) all three components are good?

Solution

Selecting a component from each box is an independent event. Let us call selecting a component from the respective boxes as events A, B and C.

(a) Here we let P represent the probability of selecting a *faulty* component.

$$P(A \text{ AND } B \text{ AND } C) = P(A) \times P(B) \times P(C)$$
$$= 0.06 \times 0.08 \times 0.05$$
$$P(A \text{ AND } B \text{ AND } C) = 0.000240$$

In other words the probability of selecting three faulty components is 0.024 per cent or 0.000240 as a decimal.

(b) Here we let P represent the probability of selecting a *good* component

$$P(A \text{ AND } B \text{ AND } C) = P(A) \times P(B) \times P(C)$$
$$= 0.94 \times 0.92 \times 0.95$$
$$P(A \text{ AND } B \text{ AND } C) = 0.82156$$

In other words the probability of selecting three good components is 82.156 per cent or 0.82156 as a decimal.

SELF-CHECK QUESTIONS

5.6 A product is made up of five independently produced components. The probability of any one component being defective is 0.03. What is the probability of all five components being 'good' (i.e. of none being defective)?

5.7 In a given batch of seeds, the probability of a seed germinating is 0.95. The probability of a blue flower coming from any one of the germinating seeds is 0.4. What is the probability that any one seed will result in a blue flower?

5.8 Two school friends always do their best at homework, but at mathematics the probability that Lee gets problems right is 3/4, while the probability that Conrad gets the problems right is 9/10. What is the probability that

(a) both get a problem correct
(b) both get a problem wrong
(c) neither gets a problem wrong?

Note: **Answers can be found on p. 339.**

5.5 Conditional probability

Here we look at situations where the probability of an event occurring depends in part on whether or not other events have already occurred. We call this **conditional probability**.

Some events are neither mutually exclusive nor independent. However, if they do happen, they influence the probability of other, subsequent, events happening. For example, if we purchase a cheaper, lower quality machine tool, then the probability of its requiring repair over the next year will be higher than if we had selected a more expensive, higher quality machine tool. In other words we are looking at *dependent* events involving situations of conditional probability.

It will help to become familiar with some terminology:

P(A) = Probability of event A
P(B|A) = Probability of event B, given that event A has already occurred.
P(C|AB) = Probability of event C, given that events A and B have already occurred, etc.

The contrast between *independent* events (AND rule) and *dependent* events (conditional probability) can be illustrated in the following example involving drawing coloured balls from a bag. When we *replace* the ball selected on the first draw before we draw the second ball, then the second draw is *independent* of the first. However when we *do not replace* the first ball before we draw the second ball, then the second draw is *dependent* on the first.

.3 A bag contains 40 balls, of which 16 are red and 24 are blue. Three balls are drawn separately from the bag. What is the probability of drawing three red balls:

(a) when each ball drawn is *replaced* before the next draw?
(b) when each ball drawn is *not replaced* before the next draw?

Let A = getting a red ball on first draw
Let B = getting a red ball on second draw
Let C = getting a red ball on third draw

$$P(ABC) = P(A) \times P(B) \times P(C)$$

$$\text{where } P(A) = P(B) = P(C) = \frac{16}{40} = 0.4$$

$$\text{So } P(ABC) = 0.4 \times 0.4 \times 0.4 = 0.064$$

In other words, with replacement, there is a 6.4 per cent or 0.064 probability of selecting three successive red balls from the bag.

(b) Here we do *not* have replacement, so the probabilities of selecting subsequent red balls are *dependent* on whether a previous red ball has been selected.

$$P(ABC) = \frac{16}{40} \times \frac{15}{39} \times \frac{14}{38} = \frac{3{,}360}{59{,}280} = 0.057 \text{ (to 3 d.p.)}$$

In other words, without replacement, there is a 5.7 per cent or 0.057 probability of selecting three successive red balls from the bag.

Bayes' theorem

The idea of conditional probability can be expressed more formally using *Bayes' theorem*. This states that the probability of two *dependent events* occurring can be expressed as the probability of the first event occurring multiplied by the conditional probability of the second event occurring given that the first has already occurred (and so on for more than two events).

Where A and B are *dependent events*

$$P(A \text{ AND } B) = P(A) \times (P(B|A)$$

but $P(B \text{ AND } A) = P(B) \times P(A|B)$

i.e. $P(A) \times P(B|A) = P(B) \times P(A|B)$

We can express this result as follows:

Formulae for Bayes theorem

$$P(B|A) = \frac{P(B) \times P(A|B)}{P(A)}$$

$$\text{i.e. } P(\text{B given A}) = \frac{P(B \text{ AND } A)}{P(A)}$$

Alternatively we can express the above as:

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}$$

$$\text{i.e. } P(\text{A given B}) = \frac{P(A \text{ AND } B)}{P(B)}$$

This result is often referred to as *Bayes' theorem* and is widely used in calculations for problems involving conditional probabilities.

WORKED EXAMPLE **5.4**

40 workers in a large office can be classified as follows:

	Clerical grade	*Administrative grade*
Male	4	6
Female	22	8

(a) If a worker is selected at random from the office, what is the probability that the worker will be in the administrative grade?
(b) If the worker selected is female, what is the probability that she will be in the administrative grade?
(c) If the worker selected is male, what is the probability that he will be in the administrative grade?
(d) If the worker selected is in the administrative grade, what is the probability of the worker being male?

Solution

(a) $P(\text{admin. grade}) = \frac{14}{40} = 0.35$

(b) We can work this out as $\frac{8}{30} = 0.267$ (to 3 d.p.) since of the 30 female workers only eight are in the administrative grade.

However we can also work this out more formally using Bayes' theorem.

Via Bayes' theorem

$$(\text{admin. grade} | \text{female}) = \frac{P(\text{admin. grade AND female})}{P(\text{female})}$$

where:

$$P(\text{admin. grade AND female}) = \frac{8}{40} = 0.2$$

$$P(\text{female}) = \frac{30}{40} = 0.75$$

$$\text{So } P(\text{admin. grade} | \text{female}) = \frac{0.2}{0.75} = 0.267 \text{ (to 3 d.p.)}$$

(c) We can work this out as $\frac{6}{10} = 0.6$ since of the ten male workers six are in the administrative grade.

Via Bayes' theorem

$$P(\text{admin. grade}|\text{male}) = \frac{P(\text{admin. grade AND male})}{P(\text{male})}$$

where:

$$P(\text{admin. grade AND male}) = \frac{6}{40} = 0.15$$

$$P(\text{male}) = \frac{10}{40} = 0.25$$

$$\text{So } P(\text{admin. grade}|\text{male}) = \frac{0.15}{0.25} = 0.6$$

(d) We can work this out as $\frac{6}{14} = 0.429$ (to 3 d.p.) since of the 14 administrative grade workers six are male.

Via Bayes' theorem

$$P(\text{male}|\text{admin. grade}) = \frac{P(\text{male AND admin. grade})}{P(\text{admin. grade})}$$

where

$$P(\text{male AND admin. grade}) = \frac{6}{40} = 0.15$$

$$P(\text{admin. grade}) = \frac{14}{40} = 0.35$$

$$\text{So } P(\text{male}|\text{admin. grade}) = \frac{0.15}{0.35} = 0.429 \text{ (to 3 d.p.)}$$

5.6 Decision trees

It is often helpful to visualise a problem involving probability on a **decision tree**. When we can clearly see *all* the possibilities in a given situation it makes it much easier to calculate the relevant probabilities. Although decision trees can be used in any situation, they are particularly helpful when solving problems involving conditional probability.

WORKED EXAMPLE **5.5**

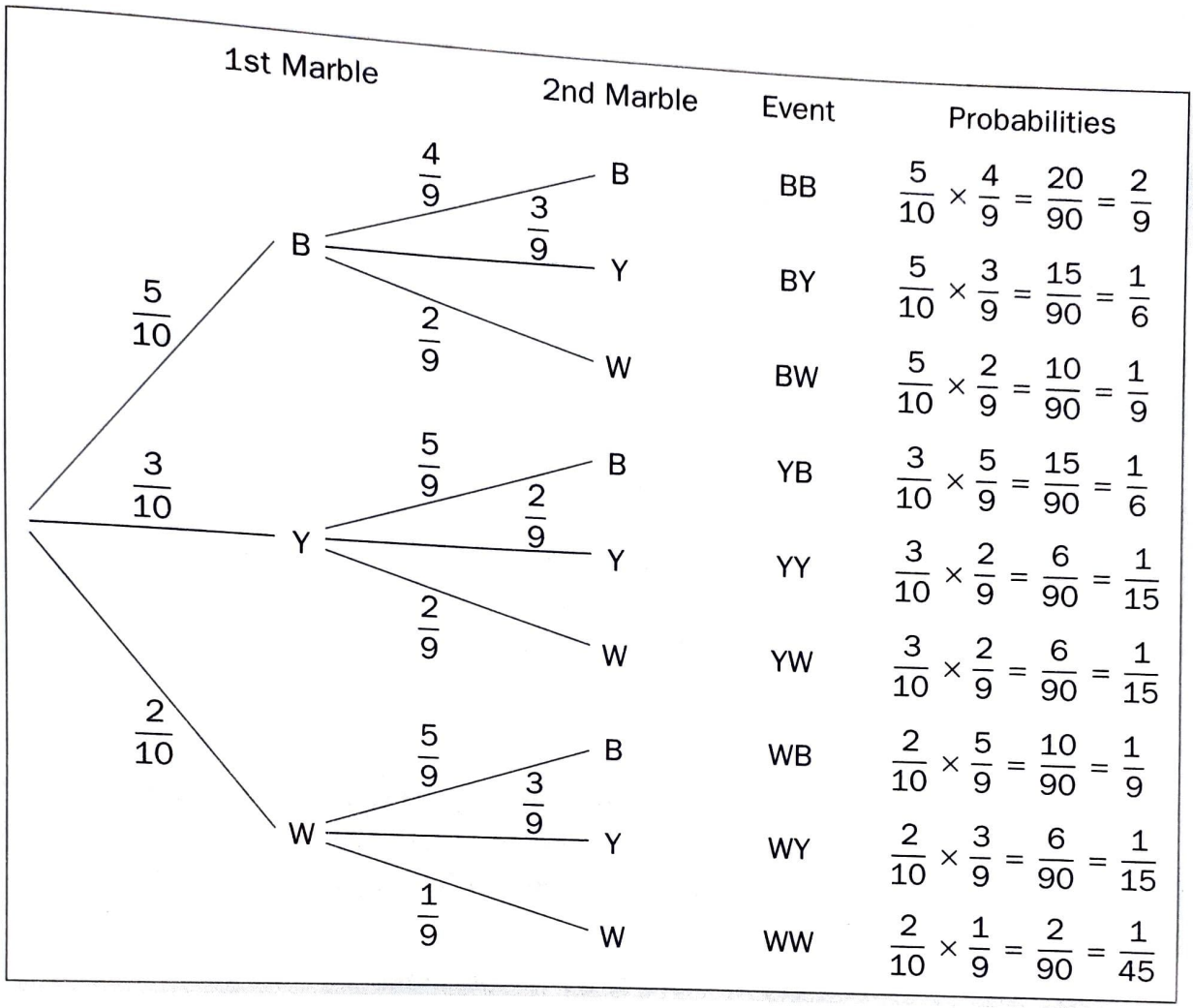
A bag contains five blue marbles, three yellow marbles and two white marbles. Two marbles are then drawn from the bag, the first marble *not* being replaced before the second marble is drawn.

What is the probability that:

(a) both marbles are the same colour
(b) both marbles are a different colour
(c) at least one marble is blue?

The decision tree (Figure 5.3) helps us see all the possibilities involved in this problem.

figure 5.3
Using a decision tree



Notice that to find the probability of an event involving any two outcomes occurring (e.g. blue marble AND blue marble, BB) we *multiply* the probabilities.

So $P(BB) = P(B) \times P(B|B)$

The probability of two blues is the probability of the first marble being blue x the probability of the second marble being blue given that the first is blue:

i.e. $P(BB) = \frac{5}{10} \times \frac{4}{9} = \frac{20}{90} = \frac{2}{9}$

However when we calculate solutions to (a), (b) or (c) below which involve either one OR more events occurring, we *add* the probabilities.

Solution

(a) Both marbles having the same colour can be achieved in three different ways, either two blues OR two yellows OR two whites.

i.e. $P(\text{both marbles same colour}) = P(BB) + P(YY) + P(WW)$

$$= \frac{2}{9} + \frac{1}{15} + \frac{1}{45} = \frac{14}{45}$$

(b) Both marbles having a different colour will have the probability of 1 *minus* the probability they both have the same colour.

i.e. P(both marbles different colour) = 1 − P(both marbles same colour)

$$= 1 - \frac{14}{45} = \frac{31}{45}$$

(c) The probability of at least one marble being blue can be achieved in five possible ways:

P(BB) OR P(BY) OR P(BW) OR P(YB) OR P(WB)

$$\frac{2}{9} + \frac{1}{6} + \frac{1}{9} + \frac{1}{6} + \frac{1}{9}$$

i.e. P(at least one marble blue) $= \frac{20 + 15 + 10 + 15 + 10}{90} = \frac{70}{90} = \frac{7}{9}$

Clearly the decision tree is a useful visual aid, helping us see more clearly how we can apply the various probability rules already learned, especially in situations involving conditional probability. As we shall see in the next section, the decision tree is also helpful in applying the idea of *expected value* to business decision making.

SELF-CHECK QUESTIONS

5.9 A box of 20 components contains six defective components. If three components are drawn consecutively from the box, what is the probability of all three components being good (non-defective):

(a) if there is replacement
(b) if there is no replacement

Use a decision tree to help solve the next problem.

5.10 Two cards are dealt out to a player from a full pack (without replacement). What is the probability of:

(a) both being a diamond
(b) one of them being a diamond
(c) neither of them being a diamond?

Note: **Answers can be found on p. 339.**

DID YOU KNOW?

The expected value of the benefits from reducing the lead content in petrol by 1 gram per gallon in the USA has been estimated at over $1 billion per annum. This calculation involved estimating the probabilities of certain health and economic benefits occurring (e.g. reduced adult blood pressure, raised children's IQs, greater fuel economy, etc.) and then placing a monetary value on each of these benefits.

5.7 Expected value

The expected value (EV) of a particular event which has different possible outcomes is a weighted average of the values (pay-offs) associated with each possible outcome. The probabilities of each outcome are used as the weights and these are *multiplied* by the respective pay-offs of each outcome.

Take, for example a situation in which there are two possible outcomes having pay-offs X_1 and X_2 with the probabilities of each outcome p_1 and p_2 respectively, then the expected value EV(X) is:

$$EV(X) = p_1X_1 + p_2X_2$$

For example, if there is a 60 per cent chance of earning £1,000 and a 40 per cent chance of earning £5,000 from an investment, then:

$$E(X) = 0.60(£1{,}000) + 0.40(£5{,}000)$$

i.e. $E(X) = £600 + £2{,}000 = £2{,}600$

More generally, using the short-hand notation of Chapter 2, the expected value (EV) of a particular course of action over n possible outcomes can be defined as:

Formulae for expected value

$$EV = \sum_{i=1}^{n} p_i \cdot X_i$$

where p_i = probability of ith outcome
X_i = value of ith outcome

and $\sum_{i=1}^{n} p_i = 1$

This last condition merely states that we have covered all possible outcomes, so the sum of their probabilities must be 1.

WORKED EXAMPLE **5.6**

A fair coin is tossed and, for a £1 stake, the gambler is promised £1.80 if the result is a head, but only £0.10 if the result is a tail. What is the expected value to the gambler of each throw?

Let p_1 = the probability of head (0.5)
p_2 = the probability of tail (0.5)
X_1 = the 'pay-off' of a head (£1.80)
X_2 = the 'pay-off' of a tail (£0.10)

Solution

$$EV = \sum_{i=1}^{2} p_i \cdot X_i$$

i.e. $EV = p_1X_1 + p_2X_2$
$EV = 0.5 \times £1.80 + 0.5 \times £0.10$
$EV = £0.90 + £0.05$
$EV = £0.95$

The gambler can therefore 'expect' to lose 5 pence on each throw of the dice, though he/she might be lucky and get more heads than tails on the early throws of the dice, and thereby win by 'quitting while they are ahead'.

On the other hand, the individual *offering* the gamble can 'expect' to make a profit (revenue – cost) of £0.05 on each throw (£1.00 – £0.95).

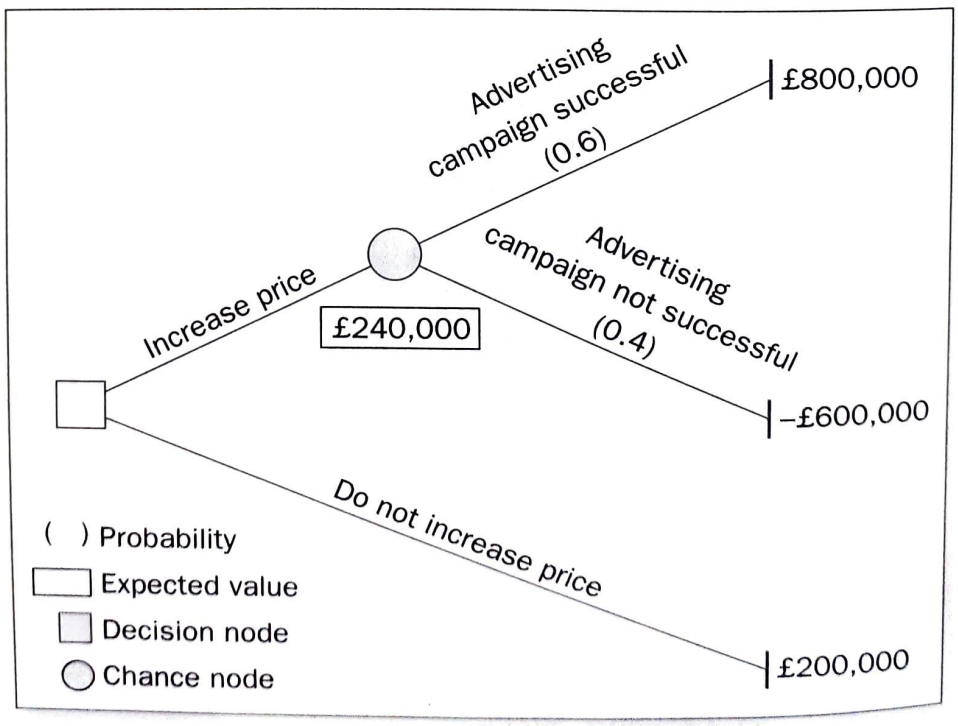
Clearly the greater the number of 'trials', e.g. throws of the dice in our example, the more likely it will be that the *theoretical probabilities* (see p. 99) will apply rather than the *experimental probabilities* (see p. 99) of a 'lucky' or 'unlucky' run.

Expected value and decision trees

We can relate this idea of expected value (EV) to the decision tree analysis already presented (p. 108). As we have seen, a decision tree is a diagram which can usefully represent a series of choices and their possible outcomes. Where the outcome is *pure chance*, then the overall probability of 1 will be divided *equally* between all the possible outcomes (e.g. 0.5 for each of two pure chance outcomes). Where the outcome is *uncertain* but where some outcomes are more likely than others, then the decision maker may assign *probabilities* to each possible outcome. The sum of these probabilities over all possible outcomes must, of course, be equal to 1.

The sequence of decisions and outcomes is represented graphically as the branches of the 'tree' (see Figure 5.4). At every point (node) that a decision must be made or an outcome must occur, the tree branches out further until all the possible outcomes have been displayed.

figure 5.4
Decision tree



- **Boxes (decision nodes)** are usually used to indicate situations where the decision maker consciously selects a particular course of action (strategy) and where the outcome of that action is 'certain'. Branches coming out of these boxes simply indicate the alternative decisions or strategies which might be taken, each of which has a probability of 1 once selected.
- **Circles (chance nodes)** are usually used to indicate situations reflecting a 'state of nature', i.e. situations whose outcomes are *not* entirely under the conscious control of the decision maker. Branches coming out of such circles show the various possibilities which might occur, together with estimates of their probability of occurrence.
- **Pay-offs** are the valuations placed at the end of particular branches emanating from chance nodes. They are the values which management allocates to that event or outcome should it actually occur, and are often denoted by a straight (vertical) line at the end of the branch.

In Figure 5.4, at the *decision node* box, the decision maker must decide between a strategy of increasing price or keeping price unchanged. If this latter strategy is pursued, only one profit outcome is possible which is evaluated at £200,000. If the former strategy (increasing price) is pursued, then the decision maker intends to support the policy change by an active advertising campaign. However the outcome of such an advertising campaign is not entirely under the control of the decision maker. He estimates that there is a 60 per cent (0.60) chance of the campaign being a success and a 40 per cent (0.40) chance of it being a failure. These two branches are therefore shown as emanating from a *chance node*, indicated by a circle. The firm estimates profits of £800,000 should the advertising campaign (allied to a price increase) be successful, but losses of £600,000 should it fail.

Backward induction

Which branch the firm should choose in order to maximise the expected profit can easily be determined. The process of solving this problem is called *backward induction*. This requires us to begin at the right-hand side of the decision tree, where the profit figures are located. The first step is to calculate the *expected profit* when the firm is situated at the chance node immediately to the left of these pay-off figures.

Because there is a 0.60 probability that the branch culminating in a profit of £800,000 will occur, and a 0.40 probability that the branch culminating in a loss of £600,000 will occur, the expected value (EV) of profit when situated at this chance node is:

$$\text{EV (profit)} = 0.60(£800{,}000) + 0.40(-£600{,}000) = £240{,}000$$

This number is written below the chance node in question to show that this is the expected profit when located at that node.

Moving further to the left along the decision tree, it is clear that the firm has a choice of two branches, one of which leads to an expected profit of £240,000, the other of which leads to an expected profit of £200,000. If the firm wants to maximise expected profit, it should choose the former branch. In other words, it should increase its price and accompany this with an advertising campaign.

WORKED EXAMPLE **5.7**

Firm A faces a choice between two policy options. One is to make a certain annual profit of £100,000 from interest on a sum of money left on deposit in a bank. The second is to invest that same sum of money in new product development (perfume).

If Firm A introduces the perfume, there is a 70 per cent probability that a major competitor, Firm B, will react by introducing a similar product and a 30 per cent probability that it will not.

Should it decide to introduce the new perfume product, Firm A can choose a high price (HP), a medium price (MP) or a low price (LP) for its new perfume product.

How much profit Firm A *estimates* that it will make depends on a number of factors.

(i) Whether or not Firm B reacts with a similar product
(ii) What strategy on price Firm A selects.
(iii) What counter strategy on price Firm B selects.

Firm A calculates the following table to represent the estimated profits (in present value terms) from the different possible outcomes. Firm A has assigned various probabilities to each possible outcome.

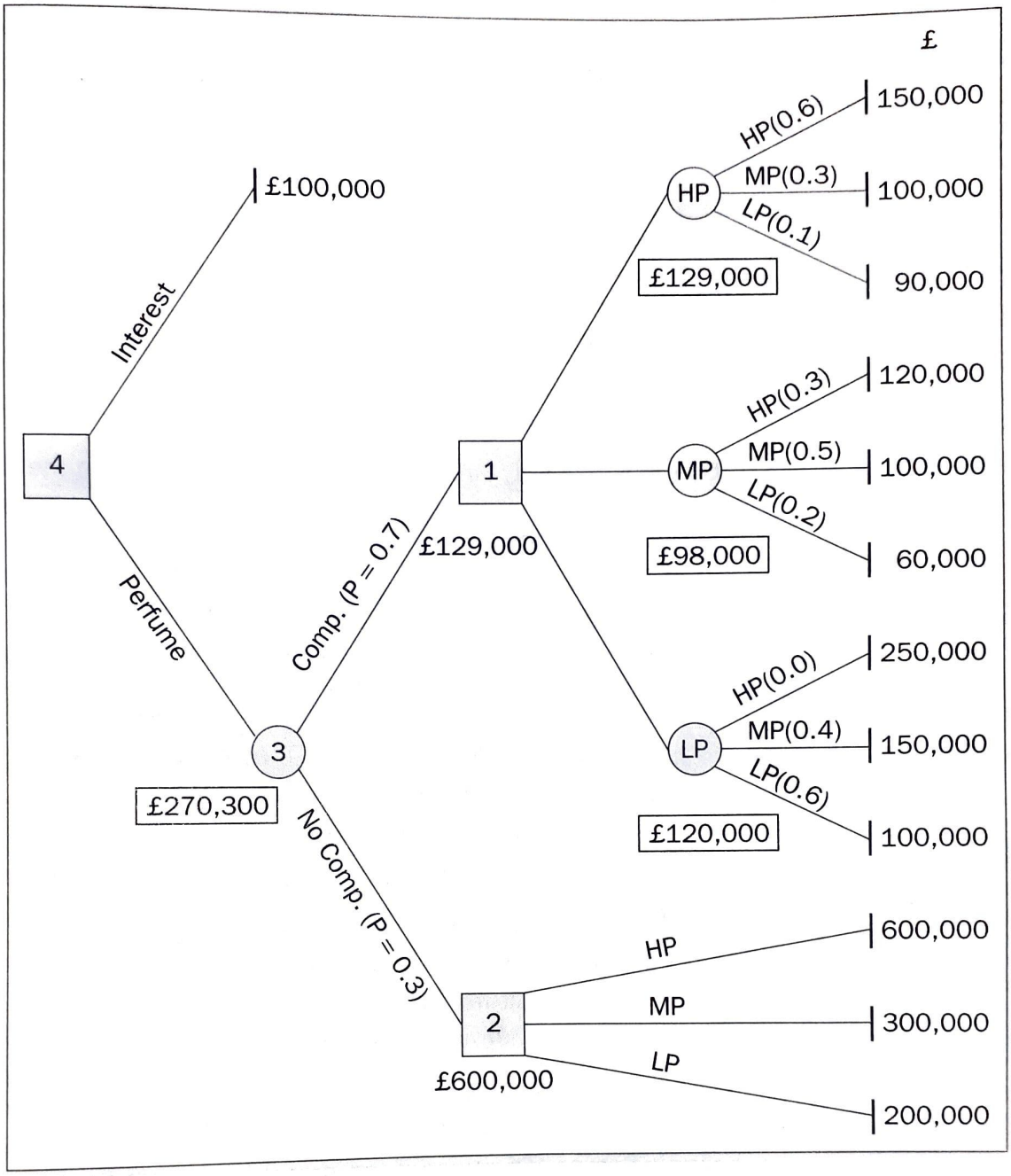
Competition	*Firm A's price strategy*	*Firm B's price response*	*Probability*	*Firm A's profit £s*
Yes	HP	HP	0.6	150,000
Yes	HP	MP	0.3	100,000
Yes	HP	LP	0.1	90,000
Yes	MP	HP	0.3	120,000
Yes	MP	MP	0.5	100,000
Yes	MP	LP	0.2	60,000
Yes	LP	HP	0.0	250,000
Yes	LP	MP	0.4	150,000
No	LP	LP	0.6	100,000
No	HP		–	600,000
No	MP		–	300,000
	LP		–	200,000

Construct a decision tree and use it to suggest whether Firm A should invest in the new perfume or leave its funds to earn interest in the bank.

Solution

Figure 5.5 presents the decision tree for this problem.

figure 5.5
Decision tree: perfume problem



Using the technique of *backward induction* we begin our solutions at the right-hand side and work leftwards, back to the initial decision node (Box 4). However we can see that there are two 'earlier' decision nodes (Boxes 1 and 2) and four chance nodes (Circles HP, MP, LP and (3). The pay-offs are listed at the end of the branches furthest to the right.

At decision box 1, Firm A must choose between a high price (HP), medium price (MP) and low price (LP) strategy. Its choice will depend on the expected value (shown underneath each chance node) of each pricing strategy, which in turn depends on the probabilities of the price responses of Firm B and the pay-offs estimated for each response.

Using our earlier expected value (EV) formula:

$$EV = \sum_{i=1}^{n} p_i X_i$$

with all values in £000s we can say:

$$EV(\text{Firm A, HP strategy}) = 0.6(150) + 0.3(100) + 0.1(90)$$
$$= 90 + 30 + 9$$

i.e. $\underline{EV\ (HP) = £129(000)}$

$$EV\ (\text{Firm A, MP strategy}) = 0.3(120) + 0.5(100) + 0.2(60)$$
$$= 36 + 50 + 12$$

i.e. $\underline{EV\ (MP) = £98(000)}$

$$EV\ (\text{Firm A, LP strategy}) = 0.0(250) + 0.4(150) + 0.6(100)$$
$$= 0 + 60 + 60$$

i.e. $\underline{EV\ (LP) = £120(000)}$

At decision node 1, Firm A will select the strategy with the highest expected profit pay-off, i.e. a higher price (HP) strategy (in the event of competition from Firm B), yielding £129,000 in expected profit.

At decision node 2, Firm A has no probabilities to contend with since the pay-offs listed are *not* associated with the possible reactions of Firm B. Here it can select the higher price (HP) strategy (in the event of no competition from Firm B) yielding £600,000 in expected profit.

However there are probabilities associated with whether or not Firm B will react with a new product of its own. Therefore we must take these into account in working out the expected value at chance node 3.

$$EV\ (\text{perfume product}) = 0.7(129) + 0.3(600)$$
$$= 90.3 + 180$$

i.e. $\underline{EV\ (\text{perfume product}) = £270.3(000) = £270,300}$

At the initial decision box (4) we can see that, based on our expectations, there will be greater profit from investing the available money in the new perfume product than from leaving the money earning interest in the bank.

An approach such as that in Worked Example 5.7 is often said to be using different *scenarios,* i.e. the evaluation of different possible outcomes for various policy initiatives. Of course if the *probabilities* were to change, for any reasons, or the expected pay-offs were to change, then any initial decision must be re-evaluated in the light of these new circumstances.

There is also the issue of the *risk* associated with any decision. For example the £100,000 earned via bank interest is likely to be much less risky than the higher expected value of £270,300 earned via new product development. We return to the ideas of investment appraisal and risk analysis in Chapter 9, p. 208.

SELF-CHECK QUESTIONS

5.11 A machine produces 35 defective items per 2,000 items produced.

(a) What is the probability of an item selected at random being defective?
(b) If 4,500 items are produced per day, how many defective items would you expect?

5.12 A company is considering introducing a new computer system for one of its northern branch offices. The company can lease a small, medium or large

computer system. The firm estimates that if the economy continues to expand over the next year the extra profits from each system will be £100,000, £150,000 and £200,000 respectively. If the economy slows down the extra profits generated will be £60,000, £20,000 and a loss of £20,000 respectively. It is estimated that the probability of continued expansion is 0.4 and of slowdown is 0.6. Construct a decision tree diagram to represent this problem and use the expected value approach to recommend one of the computer systems.

5.13 Clover design have developed a new product and are considering either test marketing it at a cost of £65,000 or abandoning it. If the project is abandoned (at any stage) the firm can recoup £20,000 from the sale of specialist equipment. The probability of a favourable result from the test market is estimated as 0.65 and an unfavourable result as 0.35. If the test market is unfavourable the project will be abandoned. A favourable test market could be followed either by abandonment of the project or by development. It is estimated that the extra profits generated will depend on the response of rival firms: if there is no response then profits will be £250,000 on the other hand, if rivals do respond and develop a similar product, then losses of £90,000 will be incurred. The probability of rivals responding is 0.3 and of not responding is 0.7. Illustrate this problem with a decision tree diagram and advise the company.

***Note:* Answers can be found on p. 340.**

5.8 Permutations and combinations

The ideas of permutations and combinations will, as we shall see, play an important role in the *probability distributions* considered in Chapter 6. It will be useful, at the outset, to become familiar with a certain type of notation often used in these types of problems.

Factorial notation

The exclamation mark of English language means something rather different here. For example 6! is the mathematical symbol for 6 factorial which means $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$. Some calculators have a key marked $x!$, and if you enter 6 and press this key you obtain 720.

Use your calculator to show that $7! = 5{,}040$.

Note: $1! = 1$ and $0! = 1$.

Permutations

Suppose you are given ten desirable qualities for a new product and are asked to place 4 of these in order of merit, i.e. first, second, third, fourth, in how many different ways can this be done?

Clearly the ordering of the four qualities selected *does* matter here. We are therefore seeking the number of possible *permutations*. The first place can in fact be chosen in ten different ways; having chosen the first place, there are nine possible places left, so that the second place can be chosen in nine different ways. Similarly, the third place can be chosen in eight different ways and the fourth place in seven different ways. The total number of ways is $10 \times 9 \times 8 \times 7 = 5{,}040$.

This may be written using factorial notation

$$10 \times 9 \times 8 \times 7 = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = \frac{10!}{6!} = 5{,}040$$

Formula for permutations

In general, if there are n items and r are to be placed *in order*, the number of different ways (*permutations*) in which this can be done is:

$${}_nP_r = \frac{n!}{(n-r)!}$$

The example above would be written

$${}_{10}P_4 \text{ i.e. } \frac{10!}{(10-4)!} = \frac{10!}{6!} = 5{,}040$$

In other words 5,040 different permutations are possible when choosing four out of ten items and then ordering them accordingly.

Combinations

Here we are *not* interested in the ordering of the items selected.

Consider again the example used for permutations. Suppose we wish to select four desirable qualities of a new product out of the ten desirable qualities given (*irrespective of the order of merit*). The number of ways of doing this is equal to the number of permutations, 5,040, *divided* by the number of ways of placing the four selected items in order of merit. In fact we can arrange the four selected items in $4 \times 3 \times 2 \times 1 = 4! = 24$ different ways. The result of 5,040 divided by 24 is 210, and this is called the number of *combinations*.

Using factorial notation, the number of combinations is equal to:

$$\frac{10!}{6! \times 4!} = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(6 \times 5 \times 4 \times 3 \times 2 \times 1) \times (4 \times 3 \times 2 \times 1)} = \frac{3{,}628{,}800}{720 \times 24} = 210$$

Formula for combinations

In general, if there are n items and r are to be selected irrespective of order, the number of ways (*combinations*) in which this can be done is:

$${}_nC_r = \frac{n!}{(n-r)! \times r!}$$

It is useful to find ${}_nC_r$ for $r = 0, 1, 2, 3 \ldots$

$$r = 0: \; {}_nC_0 = \frac{n!}{n! \times 0!} = 1$$

$$r = 1: \; {}_nC_1 = \frac{n!}{(n-1)! \times 1!} = n$$

$$r = 2: \; {}_nC_2 = \frac{n!}{(n-2)! \times 1!} = \frac{n(n-1)}{2 \times 1}$$

$$r = 3: \; {}_nC_3 = \frac{n!}{(n-3)! \times 3!} = \frac{n(n-1)(n-2)}{3 \times 2 \times 1}$$

The values

$$1, \; n, \; \frac{n(n-1)}{2 \times 1}, \; \frac{n(n-1)(n-2)}{3 \times 2 \times 1}, \; \frac{n(n-1)(n-2)(n-3)}{4 \times 3 \times 2 \times 1}, \ldots$$

are called *binomial coefficients*. We need these in the next chapter when we consider Binomial and Poisson probability distributions.

SELF-CHECK QUESTIONS

5.14 How many ways can we select three elements from a set of eight elements:

(a) where the selected elements are ranked in order?
(b) where the selected elements are not ranked in order?

5.15 A batch of 20 video recorders contains four defectives. A sample of five videos is to be selected from the batch.

(a) How many possible samples are there?
(b) If all the samples were to contain only non-defective videos, how many ways can this occur?

5.16 Five students (A, B, C, D and E) have been elected to the student union.

(a) If three students are to be selected for a committee, how many different committees are possible?
(b) If the three members are to become chairperson, treasurer and secretary, how many different line ups from the five students are possible?

***Note:* Answers can be found on p. 341.**

REVIEW QUESTIONS

5.1 An insurance company estimates that, in a large city, one home in every 50 is broken into each year. The company has recently sold 1,600 home insurance policies in the city.

(a) How many claims can the company expect in a one year period?
(b) If the typical claim is for £750 what is the minimum (ignore all other costs) the company must charge for each policy in order to break even?

5.2 *The probability that a record shop will sell at most eight copies of a particular CD each day is 0.60 and the probability that it will sell from nine to 12 copies of the CD each day is 0.35. What is the probability of:*

(a) at least nine sales;
(b) at most 12 sales;
(c) more than 12 sales?

5.3 *A sample of four items is selected from a batch that contains eight defective items and 24 without defects. What is the probability that:*

(a) all four items selected will be defective (assume replacement);
(b) all four items are not defective (again assume replacement)?
(c) How would your answers to (a) and (b) alter if there was no replacement?

5.4 *Two hundred first year business studies students enrol at a university. Seventy are male and choose the marketing stream, 16 are male and do not choose marketing, 90 are female and choose marketing, 24 are female and do not choose marketing. If a student is selected at random what is the probability that the student will:*

(a) be male;
(b) be female;
(c) be in the marketing stream;
(d) be male and study marketing?
(e) If a student is female, find the probability that she is not studying marketing.

5.5 Records show that for every 100 items produced in a factory during the day shift, two are defective, and for every 100 produced during the night shift, four are defective. If, during a 24 hour period, 2,000 items are produced during the day and 800 during the night, what is the probability that an item picked at random from the 2,800 produced during the 24 hour period:

(a) was produced in the day shift and was defective;
(b) was produced in the night shift and was defective;
(c) is defective irrespective of the shift?
(d) If a selected item is defective, what is the probability that it came from the day shift?
(e) If a selected item is defective, what is the probability that it came from the night shift?

5.6 A market research company has studied the quality of after-sales service provided by 100 electrical retailers in a region. The findings are summarised below:

	Good service	*Poor service*
High street chain	40	24
Independent retailer	26	10

(a) If a retailer is selected at random, what is the probability that the retailer gives good service?
(b) If a retailer is selected at random, what is the probability that the one selected is independent and gives good service?
(c) What is the probability that a retailer that is part of a high street chain will give poor service?

5.7 A sales representative finds that the probability of making a sale on the first visit to a new client is 0.5. On the second visit the probability of making a sale is 0.6 if a sale was made on the first visit and 0.4 if no sale was made on the first visit. Use a tree diagram to work out the probability of just one sale resulting from the two visits.

5.8 An employer is attempting to estimate the proportion of employees who drink alcohol regularly during their lunch break. In an attempt to ensure honest answers the employer asks each employee to secretly toss a coin.

If a head is obtained then the employee answers the question, 'were you born in an odd numbered year?'

If a tail is obtained then the employee answers the question, 'do you drink regularly at lunch time?'

The percentage of employees answering yes to either question was 37 per cent. Use a tree diagram to estimate the percentage of regular lunch time drinkers.
(N.B. let the probability of a lunch time drinker be p)

5.9 A company is contracted to finish a £100,000 project by 31 December. If it does not complete on time a penalty of £8,000 per month or part of a month is incurred. The company estimates that if it continues alone there is only a 40 per cent chance of completing on time and that the project might be one, two, three or even four months late with equal probability.

Subcontractors can be hired by the firm at a cost of £18,000. If the subcontractors are hired then the probability that the company completes on time is doubled. If the project is still late it will only now be one or two month late with equal probability.

Draw a decision tree to represent the problem of whether the company should continue alone or use subcontractors and analyse using the expected value technique.

5.10 In a lottery, a player wins first prize by selecting the correct six number combination from a selection of six different numbers drawn from the range 1 to 49. How many different possible combinations are there?

5.11 An office safe has buttons numbered 0–9 inclusive. In order to open the safe three different numbers have to be pressed in sequence. What is the probability that someone randomly pressing three buttons will open the safe?

5.12 Liker Airways has 15 video programmes available to show during its trans-atlantic flights. If six videos are to be shown during the flights:

(a) If the order is unimportant, how many different combinations are there?

(b) If 900 of the combinations of six programmes are thought to be incompatible, what is the probability that a selection of six taken at random will be compatible?

5.13 A textile designer has chosen to use eight different colours in this year's fabric designs. Production constraints mean that only five different colours can be used in each design.

How many different colour groupings can be advertised? (Assume all fabrics have five different colours.)

Answers to Review questions are to be found on pp. 415–419 at the back of the book.

Further study and data

Texts

Bancroft, G. and O'Sullivan, G. (1993), *Quantitative methods for accounting and business studies*, 3rd edn, McGraw Hill, chapters 12 and 13.

Curwin, J. and Slater, R. (1996), *Quantitative methods for business decisions*, 4th edn, International Thompson Business Press, chapter 10.

Lawson, M., Hubbard, S. and Pugh, P. (1995), *Maths and statistics for business*, Addison Wesley Longman, chapter 4.

Morris, C. (1999), *Quantitative approaches in business studies*, 5th edn, Pitman, chapter 8.

Oakshott, L. (1996), *Essential elements of business statistics*, BPP, chapter 4.

Swift, L. (2001), *Quantitative methods for business, management and finance*, Palgrave Publishers, parts P1–P4.

Thomas, R. (1997), *Quantitative methods for business studies*, Prentice Hall, chapter 2.

Waters, D. (1997), *Quantitative methods for business*, 2nd edn, Addison Wesley Longman, chapter 13.

Wisniewski, M. with Stead, R. (1996), *Foundation quantitative methods for business*, Pitman, chapter 8.

Sources of information and data

See the list the end of Chapter 1 (p. 25).

On-line exercises

Check the web site *www.booksites.net/burton* to find extra questions, Spreadsheet and on-line exercises for this chapter.

CHAPTER 6

Probability distributions

Objectives

When you have read this chapter you should be able to:

- understand the main properties of a Normal distribution;
- work out the probabilities of particular events occurring when these follow a *Normal* distribution;
- know that a 'confidence' interval is a range of values within which a particular value might lie;
- use and interpret the probability (Z) tables associated with the *Normal* distribution;
- be familiar with the *Binomial* and *Poisson* distributions, and be able to work out the probabilities of events occurring when these follow such distributions;
- be able to identify when and how to use the *Normal approximation* to the Binomial and Poisson distributions.

Introduction

Chapter 5 looked at the idea of probabilities and their application to business situations. In this chapter we look at the idea of a **probability distribution**. One of the most important probability distributions you will encounter in any quantitative methods course is the **Normal distribution** and its associated table of probabilities. We considered some of the properties of the Normal distribution in Chapter 2. Here we note how these properties can be used to develop the so called **Z-tables**, from which we can derive estimates of probability. The **Binomial** and **Poisson** probability distributions are also introduced and related to the work on probability in Chapter 5. Other probability distributions, such as the **Student-t** and **Chi-squared** distributions, are considered in Chapter 7.

Answers to 'Self-check questions', responses to each 'Pause for thought' and answers to the 'Review questions' can be found at the end of the book.

6.1 Normal distribution

We noted in Chapter 2 (p. 35) that the **Normal distribution** is perfectly symmetrical ('bell' shaped), with all three measures of central location, namely arithmetic mean, median and mode, being the same value for some variable X_i. If we call this arithmetic mean value the Greek letter μ ('mew'), then we know that since this is also the median value, then 50 per cent of the distribution will lie either side of μ. We can of course express such percentages as *decimals*, with the whole distribution underneath the Normal curve expressed as 1 and with 0.5 of the distribution then lying either side of μ.

Indeed the symmetry of any Normal distribution around μ will allow us to calculate probabilities involving the distribution of the variable X_i around μ. We can do this by calculating the **standard deviation** (σ) for the variable X_i.

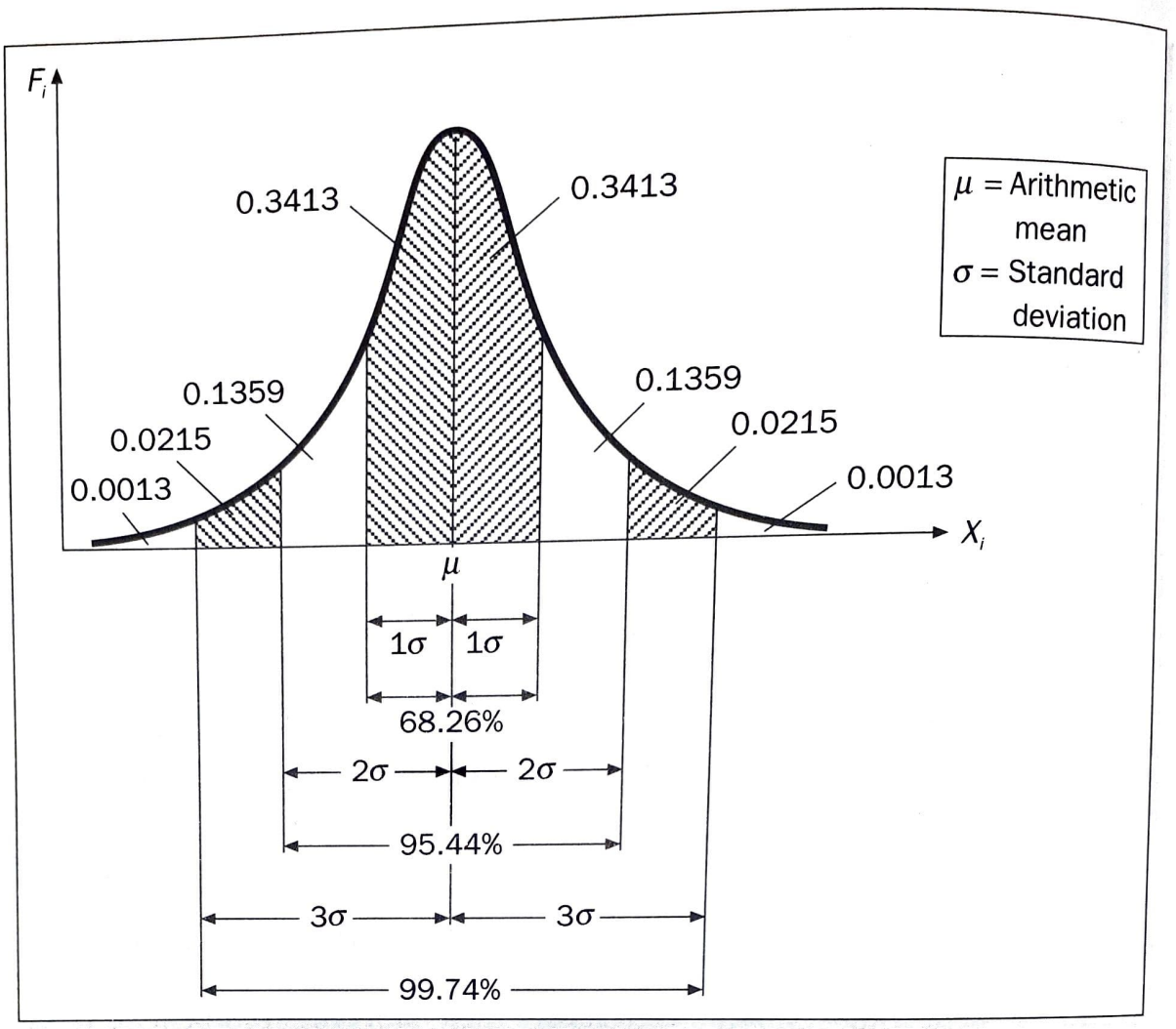


figure 6.1
Areas and probabilities under the Normal distribution

This measure of dispersion (see Chapter 2) can then be used to calculate probabilities for any Normal distribution. For example, as shown in Figure 6.1, we know that 68.26 per cent of all the observations will lie within one standard deviation of the mean; 95.44 per cent of the observations will lie within two standard deviations of the mean; and 99.74 per cent of the observations will lie within three standard deviations of the mean; etc.

PAUSE FOR THOUGHT 6.1 *Can you suggest three different types of distribution you might expect to be Normal or close to Normal?*

Confidence interval

We use the term **confidence intervals** to refer to those ranges of values from the arithmetic mean (μ) expressed in terms of the standard deviation (σ), within which we can expect a particular value of the variable X_i to lie. We return to consider confidence intervals further in Chapter 7 (p. 152).

6.2 Standard normal distribution

The calculation of probabilities under any Normal distribution can be taken a step further by transforming the values of the variable X_i into the numbers of

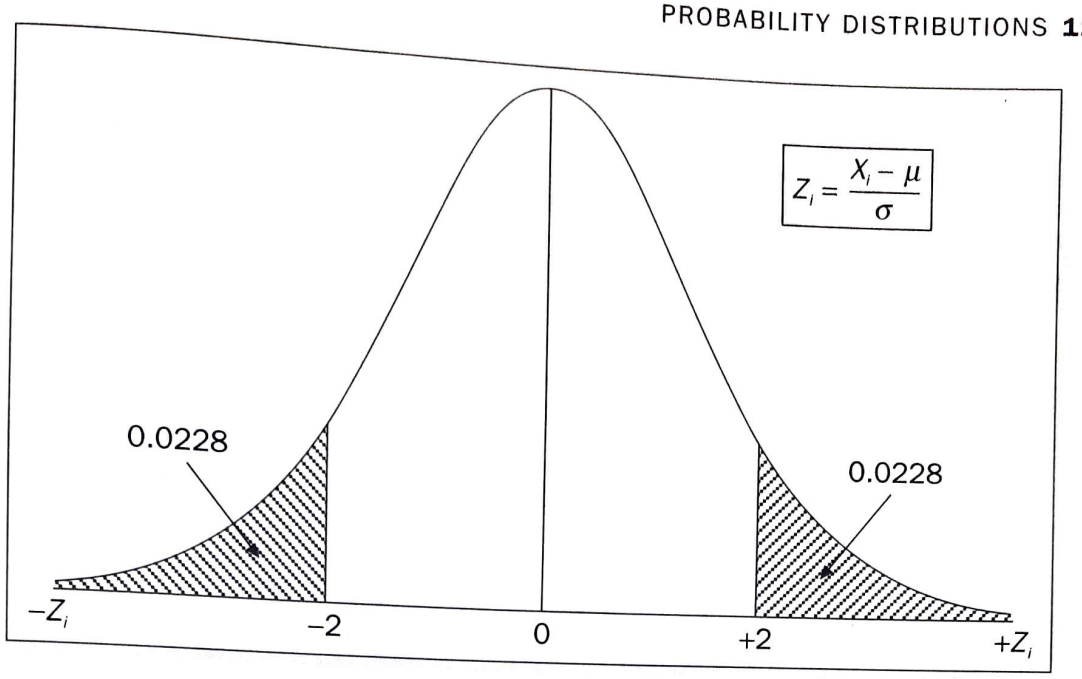


figure 6.2
The standard Normal (Z) distribution: mean 0, standard deviation 1

standard deviations (σ) from the mean (μ). The resulting distribution is often referred to as the **standard Normal distribution** and the transformation which brings it about is often called the **Z score** or **Z statistic**:

$$Z_i = \frac{X_i - \mu}{\sigma} = \frac{\text{Value of variable} - \text{Arithmetic mean}}{\text{Standard deviation}}$$

In other words the Z statistic simply calculates the number of standard deviations from the arithmetic mean. If the distribution of the variable X_i is Normal, then the distribution of the Z statistic will also be Normal, as shown in Figure 6.2.

A table giving the area shaded in Figure 6.2 for any value of the Z statistic is presented in full in Appendix 1 (p. 314). For example, for a Z score of +2.0 we will have 0.0228 (2.28 per cent) of the distribution to the *right* of that Z score.

Of course, since the Z statistic is normally distributed, a Z score of –2.0 will give us an equivalent area, but this time to the *left* of that Z score.

In other words, if the particular value for X_i is two standard deviations above (or below) the mean for a normally distributed variable, then we can expect 0.0228 (2.28 per cent) of the distribution to be above (or below) that value.

If Z = +, then area in right-hand tail
If Z = –, then area in left-hand tail

Note that the mean of the standard Normal (Z) distribution is a Z score of 0. In other words when the value of X_i is the mean (μ) itself, then we are zero standard deviations from the mean, so the Z score is 0. It therefore follows that for $Z = 0$, 0.5 (50 per cent) of the distribution will lie to the right, and 0.5 (50 per cent) to the left.

To help you become familiar with using the full Z table (p. 314), a small section of that table is shown below:

Z	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*	*0.06*	*0.07*	*0.08*	*0.09*
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681

For example, if the Z score is found to be +1.15, then go *down* the Z column to the 1.1 row and *across* the 1.1 row to the position underneath the 0.05 column. We read off the value 0.1251, which tells us that 0.1251 or 12.51 per cent of the distribution is in the *right-hand tail*. Of course had the Z score been –1.15, then 0.1251 or 12.51 per cent of the distribution would have been in the *left-hand tail*.

Check that for $Z = +1.38$ you would read 0.0838 (8.38 per cent) in the right-hand tail, and for $Z = -1.42$ you would read 0.0778 (7.78 per cent) in the left-hand tail.

WORKED EXAMPLE **6.1**

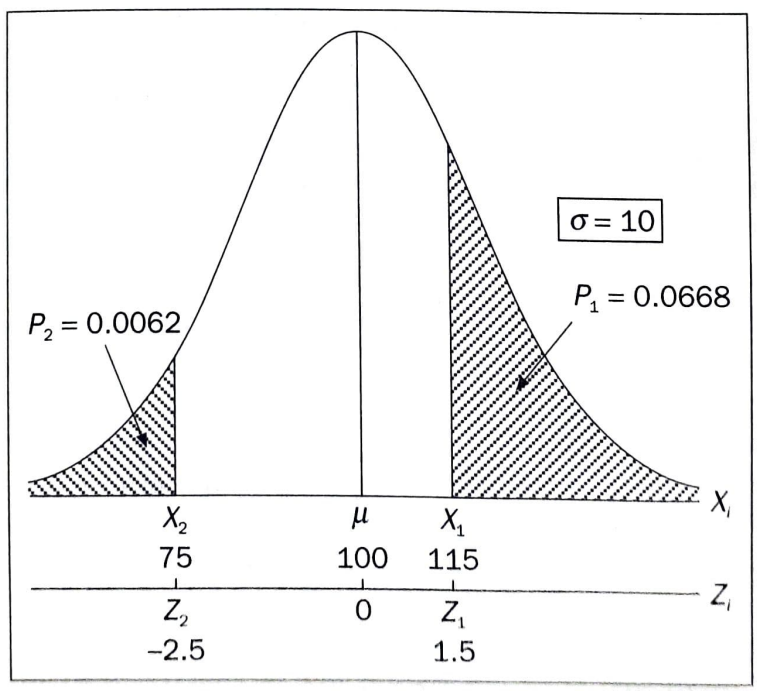
Suppose a variable X_i has a Normal distribution with a mean of 100 and a standard deviation of 10. Suppose that we want to find the probability of X_i being:

(a) 115 or more
(b) 75 or less
(c) between 75 and 115

Solution

We can use Figure 6.3 to help solve this problem. Note that we show both the variable X_i and its associated Z scores, Z_i on the same diagram.

figure 6.3



(a) For $X_1 = 115$, $Z_1 = \dfrac{X_1 - \mu}{\sigma} = \dfrac{115 - 100}{10} = +1.5$

From the Z tables (p. 314), we find that $Z_1 = +1.5$ gives 0.0668 as the area in the right-hand tail.

i.e. $Z_1 = +1.5$
$P_1 = 0.0668$
$P_1 = 6.68\%$

There is a 6.68 per cent chance of X_i being 115 or more.

(b) For $X_2 = 75$, $Z_2 = \dfrac{X_2 - \mu}{\sigma} = \dfrac{75 - 100}{10} = -2.5$

From the Z tables (p. 314), we find that $Z_2 = -2.5$ gives 0.0062 as the area in the left-hand tail.

i.e. $Z_2 = -2.5$
$P_2 = 0.0062$
$P_2 = 0.62\%$

There is a 0.62 per cent (i.e. less than 1 per cent chance) of X_i being 75 or less.

(c) The whole area beneath the Z curve is 1.00, so the area between $X_2 = 75$ and $X_1 = 115$ is $1.00 - (0.0062 + 0.0668) = 1 - 0.0730 = 0.9270$.

In other words there is a 92.7 per cent chance of X_i being between 75 and 115.

In solving problems of this kind it is always helpful to draw a sketch diagram to indicate what has been given in the question and what needs to be found, as in Figure 6.3 above. Sometimes we will need to make use of the fact that the area to each side of the mean under the normal curve is 0.50 (50 per cent). You will see in Worked Example 6.2 that sometimes we must subtract the area in the respective tails of the distribution from 0.50 in order to obtain the area required. Note also that whenever we seek to find an area which extends across the mean (μ), then it is usually helpful to solve for each side of the mean separately before combining the result.

WORKED EXAMPLE **6.2**

The attendance at a night club is thought to be normally distributed, with a mean of 80 persons and a standard deviation of 12 persons. What is the probability that on any given night the attendance is:

(a) 74 persons or more
(b) 83 persons or less
(c) between 70 and 85 persons?

Solution

Figure 6.4 represents the areas we require in each case.

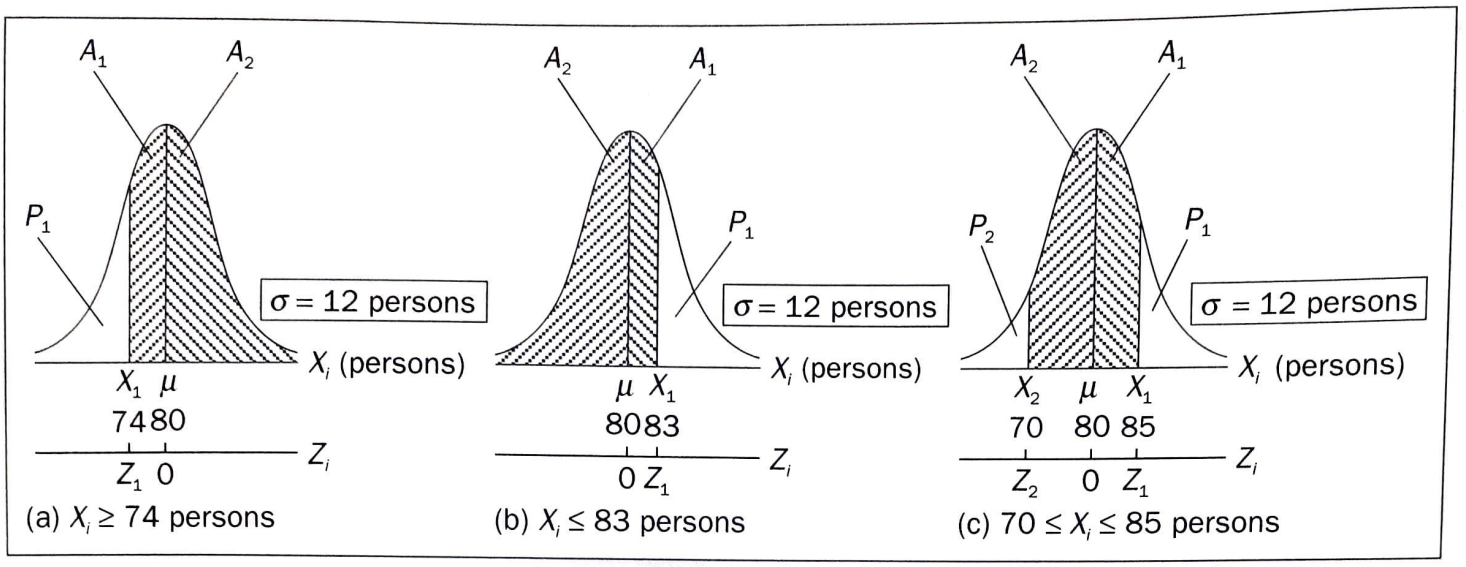


figure 6.4
The areas required are shaded (A_1 and A_2): the areas given in the *Z* tables are unshaded (P_1 and P_2)

Remember that when we calculate the *Z* score, the *Z* tables (p. 314) give us the area in the *tails* of the distribution.

(a) To find $X_i \geq 74$ means we require the shaded areas A_1 and A_2 in Figure 6.4(a).

$$\text{For } X_1 = 74,\ Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{74 - 80}{12} = -0.5$$

$Z_1 = -0.5$
$P_1 = 0.3085$
$A_1 = 0.5000 - 0.3085 = 0.1915$
$A_2 = 0.5000$
$A_1 + A_2 = 0.6915$

There is a 0.6915 (69.15 per cent) probability that the attendance will be 74 persons or more.

(b) To find $X_i \leq 83$ means we require the shaded areas A_1 and A_2 in Figure 6.4(b).

$$\text{For } X_1 = 83,\ Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{83 - 80}{12} = +0.25$$

$Z_1 = +0.25$
$P_1 = 0.4013$
$A_1 = 0.5000 - 0.4013 = 0.0987$
$A_2 = 0.5000$
$A_1 + A_2 = 0.5987$

There is a 0.5987 (59.87 per cent) probability that the attendance will be 83 persons or less.

(c) To find $70 \leq X_i \leq 85$ persons we required the shaded areas A_1 and A_2 in Figure 6.4(c).

$$\text{For } X_1 = 85,\ Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{85 - 80}{12} = \frac{+5}{12} = +0.42 \text{ (rounded)}$$

$Z_1 = +0.42$
$P_1 = 0.3372$
$A_1 = 0.5000 - 0.3372 = 0.1628$
$A_1 = 0.1628$

For $X_2 = 70$, $Z_2 = \dfrac{X_2 - \mu}{\sigma} = \dfrac{70 - 80}{12} = \dfrac{-10}{12} = -0.83$

$Z_2 = -0.83$
$P_2 = 0.2033$
$A_2 = 0.5000 - 0.2033$
$A_2 = 0.2967$

$A_1 + A_2 = 0.4595$

There is a 0.4595 (45.95 per cent) probability that the attendance will be between 70 and 85 persons.

Finding the variable value when probability known

So far we have known the *specific values* of the variable X_i and found the probabilities with which these values might occur. Here we reverse the process. Suppose we know the *probabilities* with which specific values of the variable X_i might occur, how can we then find these values of X_i? Worked Example 6.3 shows how we go about solving such problems.

WORKED EXAMPLE **6.3**

The life of a machine component is normally distributed with mean 60 hours and standard deviation 30 minutes. If 2.5 per cent (0.025) of components last more than X_1 hours, find the value of X_1.

Solution

In a question like this, do make sure that you express all values in the *same unit*, here hours or minutes. If we use hours, then the standard deviation is 0.5 hours, as shown in Figure 6.5.

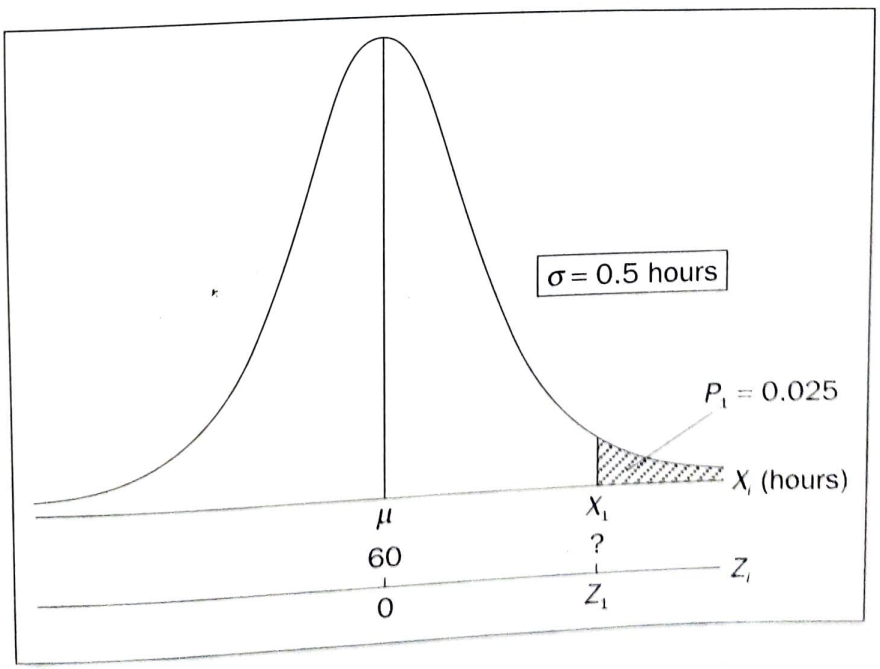


figure 6.5
Probability known, value of variable unknown

$P_1 = 0.025$

We can therefore look in the *body* of the Z tables (p. 314), to find the Z value (Z_1) which would give us the probability in the right-hand tail of 0.025.

$Z_1 = +1.96$

We know that

$$Z_1 = \frac{X_1 - \mu}{\sigma}$$

i.e. $$+1.96 = \frac{X_1 - 60}{0.5}$$

$$+1.96(0.5) + 60 = X_1$$

$$\underline{60.98 = X_1}$$

In other words we would expect 2.5 per cent (0.025) of the machine components to last more than 60.98 hours.

Note: if the left-hand tail had been involved then the calculations would have used a *negative* value for Z. You can check this yourself by finding the number of hours (X_2) for which 2.5 per cent (0.025) of the components last *less than* X_2 hours. Check that your answer is 59.02 hours and that –1.96 was used in your calculation.

SELF-CHECK QUESTIONS

6.1 Find the area beneath the (Normal) curve over the following values for the Z score. Draw diagrams in each case and use the Z tables (p. 314).

(a) $Z_1 = 0$ and $Z_2 = 0.2$
(b) $Z_1 = 0.5$ and $Z_2 = 1.0$
(c) $Z_1 = 0.95$ and $Z_2 = 2.55$
(d) $Z_1 = 1.55$ and $Z_2 = (+)$ infinity
(e) $Z_1 = -1.5$ and $Z_2 = 0$
(f) $Z_1 = -1.85$ and $Z_2 = -1.05$
(g) $Z_1 = -1.55$ and $Z_2 = (-)$ infinity

6.2 Attendants at rock concerts in a stadium are normally distributed with mean 20,000 persons and standard deviation 4,000 persons. For a future event, what is the probability:

(a) more than 28,000 persons will attend
(b) less than 14,000 persons will attend
(c) between 17,000 and 25,000 persons will attend.

6.3 Use the same data as in question 6.2.

(a) If you know that there is a 5 per cent (0.05) probability that more than a certain number of people (X_1) will attend, what is the value of X_1?
(b) If there is a 1 per cent (0.01) probability that less than X_2 people will attend, what is the value of X_2?

***Note:* Answers can be found on pp. 342–344.**

6.3 Binomial distribution

The **Binomial distribution** can be used to describe the likely outcome of events for discrete variables which:

(a) have only two possible outcomes

and

(b) are independent.

Suppose we are conducting a questionnaire. The *Binomial distribution* might be used to analyse the results if the only two responses to a question are 'yes' or 'no' *and* if the response to one question (e.g. 'yes') does not influence the likely response to any other question (i.e. 'yes' or 'no').

> Put rather more formally, the *Binomial distribution* occurs when there are n independent trials (or tests) with the probability of 'success' or 'failure' in each trial (or test) being constant.

Let p = the probability of 'success'.

Let q = the probability of 'failure'.

then $q = 1 - p$

For example, if we toss an unbiased coin ten times, we might wish to find the probability of getting four heads! Here $n = 10$, p (head) $= 0.5$, q (tail) $= 0.5$ and $q = 1 - p$

The probability of obtaining r 'successes' in 'n' trials (tests) is given by a formula which incorporates our work on combinations in Chapter 5, namely:

$$P(r) = {}_nC_r p^r q^{n-r}$$

where $r = 0, 1, 2, 3 \ldots n$

In Chapter 5 we saw that:

$${}_nC_r = \frac{n!}{(n-r)!\,r!}$$

The probability of getting exactly four heads out of ten tosses of an unbiased coin, can therefore be solved as:

$$P(4) = {}_{10}C_4 0.5^4 0.5^6$$

$$\text{now } {}_{10}C_4 = \frac{10!}{(10-4)!\,4!} = \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1} = 210$$

$$\text{So } P(4) = 210 \times (0.5)^4 \times (0.5)^6$$

$$P(4) = 210 \times 0.0625 \times 0.015625$$

$$\underline{P(4) = 0.2051}$$

In other words the probability of getting exactly four heads out of ten tosses of an unbiased coin is 0.2051 or 20.51 per cent.

It may be useful to state the formulae for finding *all* the possible probabilities of obtaining r successes in n trials.

Where $P(r) = {}_nC_r p^r q^{n-r}$

and $r = 0, 1, 2, 3, \ldots n$

then, from our knowledge of combinations (see p. 119):

$$P(0) = q^n$$

$$P(1) = npq^{n-1}$$

$$P(2) = \frac{n(n-1)}{2 \times 1} p^2 q^{n-2}$$

$$P(3) = \frac{n(n-1)(n-2)}{3 \times 2 \times 1} p^3 q^{n-3}$$

$$P(4) = \frac{n(n-1)(n-2)(n-3)}{4 \times 3 \times 2 \times 1} p^4 q^{n-4}$$

.....

$$P(n-2) = \frac{n(n-1)}{2 \times 1} p^{n-2} q^2$$

$$P(n-1) = np^{n-1} q$$

$$P(n) = p^n$$

PAUSE FOR THOUGHT 6.2 *Can you think of three different types of problem which might be solved using the Binomial distribution?*

We can, of course, use this approach to answer questions which ask for the probability of an *exact* number of 'successes' occurring, such as P(4). We can also use this approach to answer broader questions such as the probability of 'less than 4' successes occurring, which would be P(0) + P(1) + P(2) + P(3), and so on.

Cumulative Binomial distribution

An alternative to such calculations would be to use tables of the **cumulative Binomial distribution** presented in Appendix 2 (p. 315). These tables are *cumulative* in that they give the probabilities of *r or more* 'successes'. For example, if we require the probability of six or more items in a sample of ten being defective, when $p = 0.20$, then from the table we find that P(6 or more) = 0.0064. It therefore follows that for the same sample, P(5 or less) = 1 − 0.0064 = 0.9936.

WORKED EXAMPLE **6.4**

A firm produces a product and finds that 10 per cent of its output is defective. A small sample of five items is taken from the production line. Find the probability of getting each of the following numbers of defective items in the sample: 0, 1, 2, 3, 4 and 5 defective items respectively.

Solution

Here $n = 5$, $p = 0.1$, $q = 0.9$. Using the formulae listed above (p. 132), we can calculate the probabilities as being:

$P(0) = q^5 = (0.9)^5 = 0.5905$

$P(1) = npq^4 = 5 \times 0.1 \times (0.9)^4 = 0.3281$

$$P(2) = \frac{n(n-1)}{2 \times 1}p^2q^3 = \frac{5 \times 4}{2 \times 1} \times (0.1)^2 \times (0.9)^3 = 0.0729$$

$$P(3) = \frac{n(n-1)(n-2)}{3 \times 2 \times 1}p^3q^2 = \frac{5 \times 4 \times 3}{3 \times 2 \times 1} \times (0.1)^3 \times (0.9)^2 = 0.0081$$

$P(4) = np^4q = 5 \times (0.1)^4 \times 0.9 = 0.00045$

$P(5) = p^5 = (0.1)^5 = 0.00001$

So, for example, the probability of exactly two out of our five items being defective is 0.0729 or 7.29 per cent.

SELF-CHECK QUESTIONS

6.4 Evaluate the following expression using your knowledge of combinations.

(a) ${}_3C_2$ **(b)** ${}_{10}C_3$
(c) ${}_{12}C_5$ **(d)** ${}_{14}C_8$

6.5 If 10 per cent of the items produced by a manufacturer are defective, find the probability that for a sample of six items: (a) two items will be defective; (b) three items will be defective.

6.6 The probability that an invoice contains an error is 5 per cent (0.05). If an audit of ten invoices is taken, what is the probability of finding less than three incorrect invoices. Solve using the cumulative Binomial probabilities table (p. 315).

Note: **Answers can be found on pp. 344–345.**

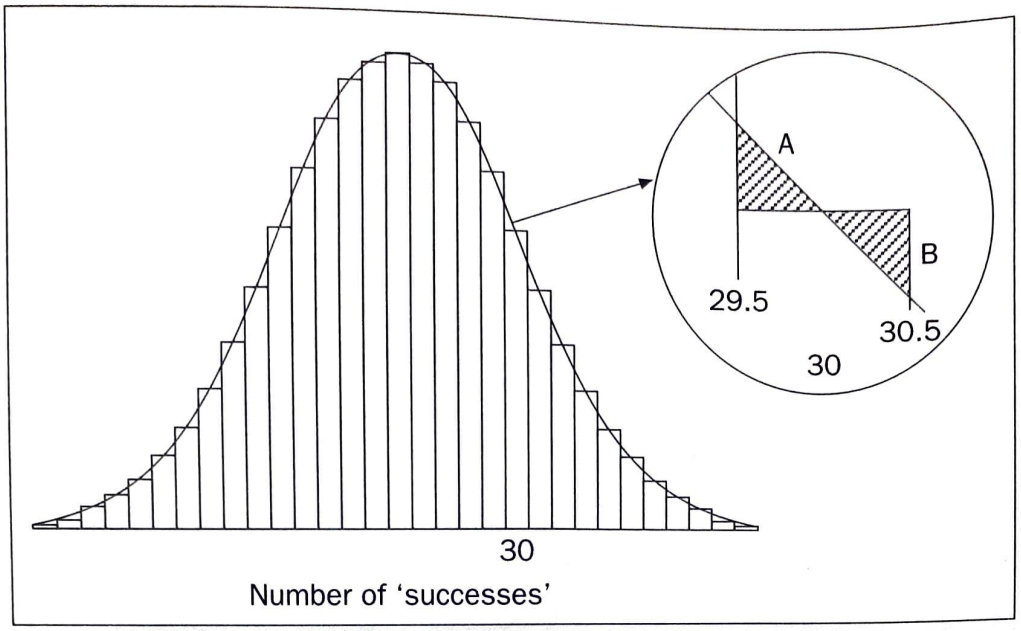
Normal approximation to the Binomial

Using the formula on p. 131 to calculate the probabilities for given numbers of successes to occur is obviously time consuming. It is therefore useful to note that if n is large and $p > 0.1$ then we can solve calculations such as those in Worked Example 6.4 using the *Normal distribution* considered at the start of this chapter. The Z statistic is calculated as before:

$$Z = \frac{X_i - \mu}{\sigma}$$

However when we are using the Normal distribution to solve, as an *approximation*, Binomial-type problems, we must express μ and σ in particular ways.

figure 6.6
Normal approximation to the Binomial distribution



For a Binomial distribution

$\mu = np$

$\sigma = \sqrt{npq}$

where p = probability of 'success'
q = probability of 'failure'
n = number of trials

Generally speaking, as a 'rule of thumb' we can use the **Normal approximation to the Binomial distribution** when $np > 5$, and $p > 0.1$

Continuity correction

Whereas the Binomial distribution represents *discrete* data, the Normal distribution represents *continuous* data. Strictly speaking, whenever we are approximating a discrete variable with a continuous one we should use a **continuity correction**.

The idea behind the continuity correction is illustrated in Figure 6.6. The curve, representing the continuous distribution, cuts through the blocks of the histogram, representing the discrete distribution.

Clearly some areas, such as A are *included* under the continuous distribution, while other areas, such as B, are *excluded* under the continuous distribution. Overall these areas will tend to cancel each other out. However, since each block of the discrete (Binomial) distribution represents a whole number, such as the number of 'successes', it can be regarded as extending from 0.5 below and above that whole number.

Thus, in Figure 6.6, the block representing 30 'successes' extends from 29.5 to 30.5. It follows that to find the area, and therefore the probability, for a series of outcomes, it will be necessary to go from 0.5 below the lowest whole number (integer) to 0.5 above the highest whole number (integer).

In Figure 6.6, to find the probability of between, say, 26 and 30 'successes', we must find the area under the normal curve from 25.5 to 30.5.

P(26 to 30) = P(25.5 to 30.5)

In Figure 6.6, to find the probability of, say, 30 or more 'successes', we must find the area under the normal curve to the right of 29.5.

P(30 or more) = P(29.5 or more)

Continuity correction

To adjust *discrete* data to *continuous* data, find the probability from 0.5 below the lowest whole number (integer) to 0.5 above the highest whole number (integer).

WORKED EXAMPLE **6.5**

Past data suggest that for every 100 telephone enquiries at the reception desk of a hotel, only some 20 bookings are made. Find the probability that from the next 100 enquiries received, the hotel will receive 30 or more bookings.

Solution

This is a Binomial situation as each enquiry leads to only one of two outcomes, booking (success) or non-booking (failure). There is an established or fixed probability ($p = 0.2$) of any enquiry leading to a booking, based on past experience.

However because n is large ($n = 100$) and p is > 0.1 ($p = 0.2$), then we can use the Normal approximation to the Binomial theorem. Note that with the continuity correction, X_1 will be 29.5 rather than 30.

$p = 0.2$

$q = 0.8$

$\mu = np = 100(0.2) = 20$

$\sigma = \sqrt{npq} = \sqrt{100(0.2)(0.8)} = 4$

$$Z_1 = \frac{X_1 - np}{\sqrt{npq}} = \frac{29.5 - 20}{4} = +2.38$$

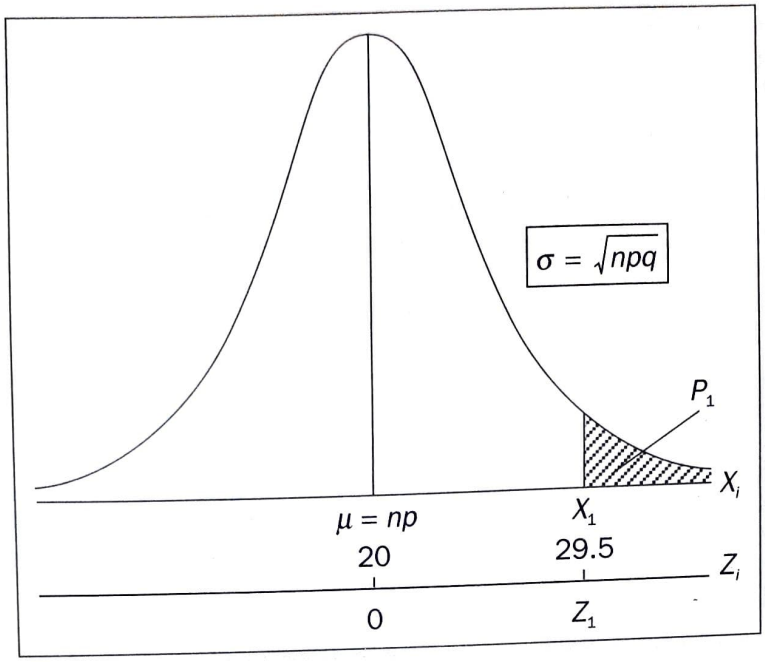
$Z_1 = +2.38$

Using the Z tables (p. 314)

$P_1 = 0.0087$

There is a 0.0087 (0.87 per cent) probability of the next 100 hotel enquiries yielding 30 or more bookings.

figure 6.7
Normal approximation to the Binomial distribution



Clearly this is a much quicker method than finding P(30), P(31) . . . P(100) using the conventional formula for the Binomial distribution presented on p. 131.

SELF-CHECK QUESTIONS

6.7 From the previous hotel booking problem calculate the probability that the next 100 enquiries will yield:

(a) 25 or more bookings **(b) 16 or more bookings**
(c) less than 14 bookings

6.8 Suppose that 10 per cent of all houses in a district have burglar alarms. A random sample of 121 houses is selected; use the Normal approximation to find the probability that:

(a) Exactly 14 of the houses have alarms?
(b) Fewer than ten houses have alarms?
(c) From seven to 11 inclusive of the selected houses have alarms?
(d) From nine to 13 inclusive of the selected houses have alarms?

***Note:* Answers can be found on pp. 345–346.**

6.4 Poisson distribution

The **Poisson distribution** may be regarded as a special case of the Binomial distribution. As with the Binomial distribution, the Poisson distribution can be used where there are only two possible outcomes, 'success' (p) or 'failure' (q), and these events are independent. The Poisson distribution is usually used where n is very large but p is very small, and where the mean np is constant and typically < 5. As p is very small ($p < 0.1$ and often much less), then the chance of the event occurring is extremely low. The Poisson distribution is therefore typically used for unlikely events such as accidents, strikes, etc.

The Poisson distribution is also used to solve problems where events tend to occur at random, such as incoming phone calls, passenger arrivals at a terminal, etc.

Whereas the formula for solving Binomial problems uses the probabilities, for both 'success' (p) and 'failure' (q), the formula for solving Poisson problems only uses the probabilities for 'success' (p).

If μ is the mean, it is possible to show that the probability of r successes is given by the formula:

$$P(r) = \frac{e^{-\mu}\mu^r}{r!}$$

where e = exponential constant = 2.7183
μ = mean number of successes = np
n = number of trials
p = probability of 'success'
r = number of successes

If we substitute $r = 0, 1, 2, 3, 4, 5 \ldots$ in this formula we obtain the following expressions:

$$P(0) = e^{-\mu} \qquad P(1) = \mu e^{-\mu} \qquad P(2) = \frac{\mu^2 e^{-\mu}}{2 \times 1}$$

$$P(3) = \frac{\mu^3 e^{-\mu}}{3 \times 2 \times 1} \qquad P(4) = \frac{\mu^4 e^{-\mu}}{4 \times 3 \times 2 \times 1} \qquad P(5) = \frac{\mu^5 e^{-\mu}}{5 \times 4 \times 3 \times 2 \times 1}$$

In questions you are either given the mean μ or you have to find μ from the information given, which is usually data for n and p; μ is then obtained from the relationship $\mu = np$.

You have to be able to work out e raised to a negative power. Remember from your work on powers (p. 293) that e^{-3} is the same as $\frac{1}{e^3}$ so you can simply work this out using $\frac{1}{2.7183^3}$.

Alternatively, many calculators have a key marked e^x. The easiest way to find (say) e^{-3} on your calculator is to enter 3, press +/– key, press e key, and you should obtain 0.049787. If your calculator does not have an e key but has an x^y key, enter 2.7813, press x^y key, enter 3, press +/– key, then press = key; you should obtain 0.049786.

WORKED EXAMPLE 6.6

(a) In a large company with many employees, accidents occur at random at the mean rate of 3 per day. Calculate the probability that four accidents occur on a single day.

(b) There is a 0.001 probability of a faulty call on a telephone line and 2,000 calls are made each day. What is the probability of there being less than three faulty calls on that line in a day?

Both (a) and (b) can be answered using the Poisson distribution.

Solution

(a) Here accidents occur at random, and relatively few accidents occur. Although p (accidents) is unknown, $\mu = np = 3$ is known. The occurrence of non-accidents (q) is irrelevant here. We can therefore use the Poisson distribution to solve the problem.

$$P(r) = \frac{e^{-\mu}\mu^r}{r!}$$

$$P(4) = \frac{e^{-3}3^4}{4 \times 3 \times 2 \times 1}$$

$$P(4) = \frac{0.0498 \times 81}{24} = 0.1681$$

$$\underline{P(4) = 0.1681}$$

In other words there is a 0.1681 (16.81 per cent) probability of four accidents occurring on a single day.

(b) Here the probability p is very low ($p = 0.001$) and the number of trials, n, is very large ($n = 2{,}000$). The mean (μ) $= np = 2{,}000\ (0.001) = 2$ faulty calls per day. Since it is also true that $\mu = np < 5$ the Poisson distribution would again seem appropriate.

To get less than three faulty calls means P(0) + P(1) + P(2)

Using $P(r) = \frac{e^{-\mu}\mu^r}{r!}$

We get $P(0) = \frac{e^{-2}2^0}{0!} = \frac{e^{-2}1}{1} = e^{-2} = 0.1353$

$$P(1) = \frac{e^{-2}2^1}{1} = 0.2706$$

$$P(2) = \frac{e^{-2}2^2}{2 \times 1} = 0.2706$$

So the probability of getting less than three faulty calls is 0.1353 + 0.2706 + 0.2706 = 0.6765 or 67.65 per cent.

Cumulative Poisson distribution

An alternative to such lengthy calculations would be to use tables of the **cumulative Poisson distribution** presented in Appendix 3 (p. 316). These tables are *cumulative* in that they give the probabilities of *r or more* 'successes'.

For example, if we require the probability of less than three faulty calls occurring in a single day when they typically occur at random at the mean rate of two per day on a telephone line (see Worked Example 6.6(b) above), then we *adapt* r to be *three or more* for purposes of the cumulative tables. So for $\mu = 2$, $r = 3$ we read off the value 0.3233 from the cumulative tables. The probability of *less than three* faulty calls is, of course, 1 – probability of *three or more* faulty calls, i.e. 1 – 0.3233 = 0.6767. This is close to the probability found by using the

lengthier method P(0) + P(1) + P(2) for the Poisson distribution in the answer to Worked Example 6.6(b) above.

Normal approximation to the Poisson distribution

If the Poisson distribution is appropriate *except* that μ (= np) > 30 (instead of μ = < 5) then the **Normal approximation to the Poisson distribution** can be used. The Z statistic is calculated as before and the Z tables of Appendix 1 (p. 314) can be used to find probabilities in the respective tails of this distribution.

$$Z = \frac{X_i - \mu}{\sigma}$$

where $\mu = np$

$$\sigma = \sqrt{np(1-p)}$$

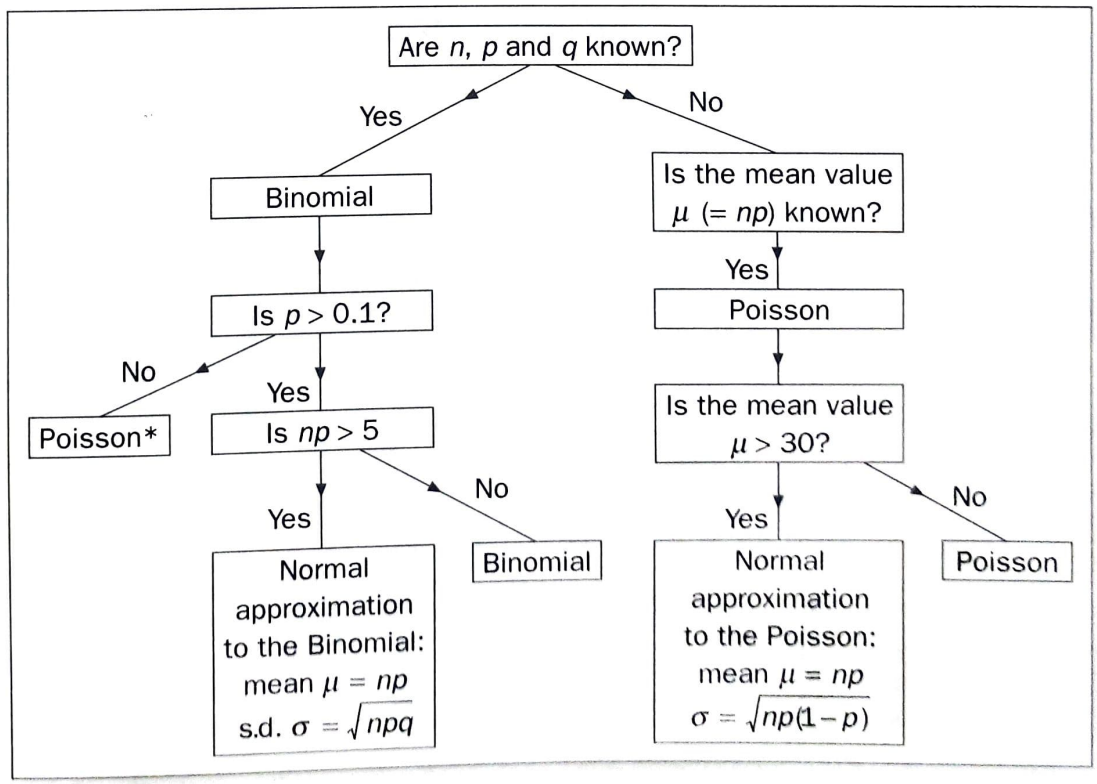
i.e. $$Z = \frac{X_i - np}{\sqrt{np(1-p)}}$$

Note: the continuity correction when adjusting discrete data to continuous data (see p. 135) will also apply here.

The following flow chart will give you some guidance on when to use the Binomial or Poisson distributions, or the Normal approximations to these distributions.

Of course the flow chart only applies to variables which are discrete (take on fixed values), which can take one of two possible outcomes (p,q) and which occur independently.

figure 6.8
Selecting the distribution to use (discrete variable, two outcomes, each independent)



Note: *Poisson is used here as an approximation to the Binomial.

SELF-CHECK QUESTIONS

6.9 Suppose $n = 1{,}000$ and $p = 0.0002$ use the Poisson distribution to find:

(a) $P(X = 0)$ (b) $P(X = 4)$
(c) $P(X > 1)$ (d) $P(X \leq 4)$
mean value $\mu = np = 1{,}000 \times 0.0002 = 0.2$

6.10 The average number of trips per family to amusement parks in the UK is Poisson distributed with a mean of 0.7 trips per year. What is the probability of randomly selecting a family and finding that:

(a) The family did not make a trip to an amusement park last year?
(b) The family took two or more trips last year?
(c) The family took three or fewer trips over a three year period?
(d) Took exactly four trips over a six year period?

***Note:* Answers can be found on p. 347.**

REVIEW QUESTIONS

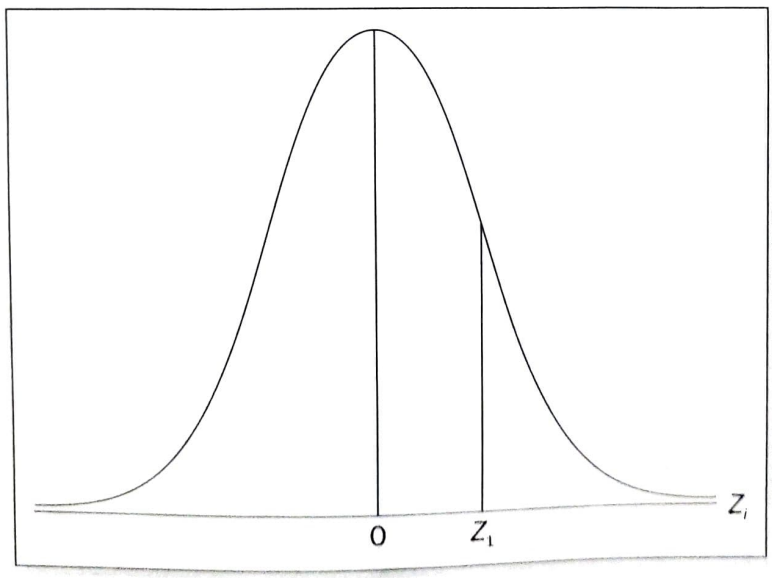
6.1 If the variable X_i follows a Normal distribution with mean = 100 and standard deviation = 5, find the probability that:

(a) $P(X_i > 107)$ (d) $P(101 < X_i < 109)$
(b) $P(X_i > 98)$ (e) $P(91 < X_i < 94)$
(c) $P(X_i < 95)$ (f) $P(93 < X_i < 106)$

6.2 Find, using the standard Normal tables (Z tables, p. 314), the value of Z_1 for each of the following:

(a) $P(Z > Z_1) = 0.3300$
(b) $P(0 < Z < Z_1) = 0.4732$
(c) $P(Z < Z_1) = 0.8212$
(d) $P(Z < Z_1) = 0.5478$

figure 6.9



6.3 The mean height of 4,000 children in a town is 130 cm with a standard deviation of 7 cm. Assuming the heights to be normally distributed, find how many children are of:

(a) height over 144 cm
(b) height less than 123 cm
(c) between 126.5 cm and 137 cm
(d) determine the upper and lower quartile heights.

6.4 The daily delivery of mail at a large city firm follows a time pattern conforming to the Normal distribution, with a mean time of arrival of 9.40 a.m. and with a standard deviation of 20 minutes.

Estimate the number of occasions during the 250 working days in the year when the mail arrives

(a) before the main gates open at 9 a.m.
(b) after the arrival of the office staff at 9.20 a.m.
(c) between 10 a.m. and 10.20 a.m.

6.5 The life of a certain type of machine component is known to be approximately normally distributed with mean 25 hours and standard deviation 30 minutes. Calculate the probability that a randomly selected component lasts for:

(a) between 24 and 24.5 hours
(b) between 24.25 and 24.75 hours
(c) more than 24.5 hours
(d) less that 24 hours 15 minutes
(e) If 28.1 per cent of the components last less than 'K' hours, find the value of K.

6.6 An engineer designs a car seat to comfortably fit women above the height of 157.5 cm. The mean height of women in the UK is 160.9 cm with a standard deviation of 6.0 cm. What proportion of women will be excluded from buying this car? (Assume the height of women is normally distributed.)

6.7 A tyre company has invented a revolutionary new product. In order to overcome consumer resistance a mileage guarantee is offered with the tyre. Road tests suggest that the mean life of the tyre is 42,000 miles, with a standard deviation of 4,000 miles. The tests suggest that tyre life is normally distributed.

(a) What percentage of tyres will last for more than 45,000 miles?
(b) What should the guaranteed mileage be if the firm wishes to replace no more than 4 per cent of tyres?

6.8 The amount of coffee that an automatic filling machine puts into nominal 100 g jars, is normally distributed and varies from jar to jar. The machine fills 50,000 jars per week and is currently set to a mean fill of 100.50 g. All jars below the nominal weight are rejected. The current reject rate is 1,500 jars per week and each reject costs the firm 80p.

(a) Find the standard deviation of the jar weights
(b) If the reliability (standard deviation) of the machine cannot be changed and the firm wishes to save £400 per week on reject costs, to what mean fill should the machine be set?

6.9 Air Inexcel has a policy of booking as many as 16 persons for an aeroplane that seats only 14. Records show that only 85 per cent of booked passengers actually arrive for the flight. Find the probability that if Inexcel books 16 people on a flight, enough seats will not be available.

6.10 A firm has agreed to accept delivery of a component shipped in batches of 1,000 if, in a random sample of ten, no more than one component is defective. What is the probability that this acceptance sampling rule would lead to a batch with 20 per cent defective items being accepted?

6.11 Suppose that 40 per cent of all students have their own personal computer. If 24 students are selected at random, find:

(a) the probability that eight will have computers;
(b) the probability that eight will have computers (using the Normal approximation to the Binomial). Compare the results with part (a).

6.12 An airline offers all 180 passengers on the London/Madrid flight a lunch choice of chicken kiev or beef in beer. As passengers choose once the flight is in progress, there is no way of knowing in advance how many will choose beef on any particular flight. The airline does not want to disappoint customers but at the same time, wants to avoid the food waste and cost of carrying excessive meals. Historical evidence suggests that 65 per cent of passengers choose chicken when offered the choice. What is the probability that, on any particular flight, the number of passengers choosing chicken will be more than 125? (use the Normal approximation to the Binomial to evaluate).

6.13 The reception desk in a hotel usually receives four guests per five minute period. Assuming that the conditions for a Poisson distribution hold, find:

(a) the probability that over a five minute period exactly four guests will arrive?
(b) the probability that two or more people will arrive?
(c) the probability that exactly six people will arrive over a ten minute period?

6.14 A certain process produces 100 m long rolls of high quality silk. In order to assess quality a 10 m sample is taken from the end of each roll and inspected for blemishes. The number of blemishes in each sample is thought to follow a Poisson distribution with an average of two blemishes per 10 m sample. Determine the probability that:

(a) there will be no blemishes observed in a 10 m sample
(b) there will be more than seven blemishes observed if a 30 m sample is taken

6.15 The proportion of defective components produced by a firm is 0.6 per cent. The components are sold in boxes of 250 and the company guarantees to replace any box containing more than three defectives. The cost of each replacement box is £4. The company is thinking about introducing a more comprehensive inspection scheme that would cost 10p per box but would eliminate all defectives. Use the Poisson approximation to the Binomial to decide whether this inspection scheme is worthwhile.

Answers to Review questions can be found on pp. 419–424.

Further study and data

Texts

Bancroft, G. and O'Sullivan, G. (1993), *Quantitative methods for accounting and business studies*, 3rd edn, McGraw Hill, chapter 14.

Curwin, J. and Slater, R. (1996), *Quantitative methods for business decisions*, 4th edn, International Thompson Business Press, chapters 11 and 12.

Lawson, M., Hubbard, S. and Pugh, P. (1995), *Maths and statistics for business*, Addison Wesley Longman, chapters 5 and 6.

Morris, C. (1999), *Quantitative approaches in business studies*, 5th edn, Pitman, chapter 9.

Oakshott, L. (1996), *Essential elements of business statistics*, BPP, chapters 5 and 6.

Swift, L. (2001), *Quantitative methods for business, management and finance*, Palgrave Publishers, parts P1–P4.

Thomas, R. (1997), *Quantitative methods for business studies*, Prentice Hall, chapter 2.

Waters, D. (1997), *Quantitative methods for business*, 2nd edn, Addison Wesley Longman, chapter 14.

Wisniewski, M. with Stead, R. (1996), *Foundation quantitative methods for business*, Pitman, chapter 8.

Sources of information and data

See the list at the end of Chapter 1 (p. 25).

On-line exercises

Check the web site *www.booksites.net/burton* to find extra questions, Spreadsheet and on-line exercises for this chapter.

CHAPTER 7

Sampling and tests of hypotheses

Objectives

When you have read this chapter you should be able to:

- appreciate the key features of the different types of sample;
- be aware of the properties of the distribution of sample means, e.g. the *standard error*;
- apply the '*Z*' score and tables to the distribution of sample means drawn from a *Normal population*;
- understand the Central Limit Theorem;
- apply the '*Z*' score and tables to the distribution of sample means drawn from a *skewed population*;
- understand and calculate *confidence intervals* for either the population mean or the sample mean;
- know when and how to use the *Student t distribution*;
- know when and how to use the *Chi-squared distribution*.

Introduction

Chapter 6 introduced the Normal distribution and the Z statistic. In this chapter we continue to apply these ideas, but in the context of **sampling**. Most firms would find it either impractical or too expensive to survey *all* their customers or to carefully examine every item that flows from their production line. Instead they usually resort to selecting a *sample* from the whole group or, as it is often called, the *population*. Of course here the term 'population' can refer to all the items under consideration as much as to all the people.

We briefly consider the *types* of sample that might be selected, and when each type might be most appropriate. We then go on to consider the **distribution of sample means** which, as we shall see, can under certain conditions be regarded as being normally distributed, thereby allowing us to use the Z table probabilities of Chapter 6. The **Central Limit Theorem** is suggesting that even when the 'population' is skewed, a random sampling procedure using large samples can still result in a Normal distribution for the sample means.

We move on to explore **confidence limits**, i.e. the ranges of values within which we can, with varying degrees of confidence, expect an observation to lie. This is used as the basis for testing various theories or **hypotheses**. When we use *large samples* ($n \geq 32$) for such tests, then the Z tables are still appropriate. However when we use *small samples* ($n < 32$), then we must use the **student t distribution**. Finally a number of other distributions used to test hypotheses are introduced such as the **Chi-squared distribution**.

The term *statistical inference* is often used to refer to the many procedures discussed in this chapter whereby attributes are ascribed to the 'population' based on the results of our sample investigations.

Answers to 'Self-check questions', responses to each 'Pause for thought' and answers to 'Review questions' can be found at the end of the book.

7.1 Types of sample

Different **types** of sample may be used, depending in part on the characteristics or attributes of the 'populations' to be sampled and on the objectives of those undertaking the sample. Here we briefly review some of the key features of the various types of sample.

Random sampling

A **random sample** is used when the intention is to give each item in the 'population' as much chance of being selected in the sample as every other item. A common way to conduct random sampling is to list all the 'members' or items in the population and then to use **random number tables** (merely a collection of random digits) or computer generated random numbers to help select the sample. Such procedures avoid various types of 'bias' which might creep into the selection of items to be included in the sample when undertaken by humans. Even the selection of numbers by individuals might be influenced by unconscious preferences for odd, even numbers, etc.

The key objective of random sampling is to obtain a sample that fairly reflects the population as a whole. Random sampling is more likely to be used when the population itself is relatively homogeneous, i.e. is composed of items of broadly the same type. Of course there may be considerable practical difficulties in actually locating and interviewing each 'member' of the population that has been randomly selected. For example, if the chosen adult is not at home or otherwise available a researcher may have to return or seek to contact that individual on another occasion – another member of the household, etc. will not be acceptable.

Suppose the random number table (see Appendix 4, p. 317) is used in seeking a random sample of, say, 40 items. This might involve linking together two columns of the five digits displayed in the table. These might begin 83635 . . . , suggesting we choose the first listed item in the first column, then 66791, and so on.

Random sampling

- ▶ each item selected has an equal chance of being drawn;
- ▶ usually adopted when the population is largely homogeneous, i.e. when it is difficult to distinguish between items;
- ▶ implementation often involves the use of computer-generated random numbers;
- ▶ selection is unbiased;
- ▶ a major drawback is that a population listing is required and the chosen items need to be located, then questioned or measured.

PAUSE FOR THOUGHT 7.1 *Can you identity three well-known products for which a random sampling procedure might be considered appropriate for testing customer response?*

The random sampling procedure will also be important should we wish to apply the **Central Limit Theorem** (see pp. 150–152).

Stratified sampling

Of course there will be times when we might prefer a *non-random* sampling procedure. Such a non-random procedure might, for example, be preferred when the population is *segmented* or *heterogeneous*, i.e. when it contains very different sub-sets of items or subjects. If, say, 40 per cent of all adults have incomes over £x and 15 per cent have incomes below £y, then we might wish our sample of adults to exactly mirror such proportions, rather than give every adult an equal chance of selection. This is the basis of **stratified sampling**. Note that the intention here is still to obtain a representative sample, but one which 'fairly' reflects a population which itself is heterogeneous. We might say that such a sample is free from 'selective bias', since the proportion of any identified attribute in the sample is merely a reflection of its contribution to the population as a whole.

Stratified sampling

- ▶ used when the population has a number of identifiable attributes;
- ▶ populations stratified in this way are known as heterogeneous;
- ▶ the composition of the sample must reflect the attributes present in the population, e.g. the proportion of low-, middle- and high-income earners;
- ▶ individuals or items *within* each stratum may still be selected randomly;
- ▶ a stratified sample is free from selective bias, since it reflects the proportions of any given attribute present in the population as a whole.

DID YOU KNOW?

It costs five times as much to recruit a new customer via marketing strategies (e.g. sample surveys of new products/designs, etc.) as it does to service and retain an existing customer.

DID YOU KNOW?

By the year 2011 it is estimated that 17 per cent of the UK population will be aged over 65 years. This 'ageing' population will mean an increased demand for those products targeted toward the elderly.

Quota sampling

The use of **quota sampling** is widespread in market research. Here the intention is often to deliberately introduce selective bias into the samples, in the sense that attributes of the members or items selected will represent the choice of the samples rather than the attributes of the population as a whole. In this sense there is no attempt to seek a representative or unbiased sample from the population. For example, if a firm sells most of its product to those with incomes over £x, then the sample may contain 80 per cent of such adults even if they only comprise, say, 40 per cent of the population. Nor is there usually any attempt to use random sampling *within* the quotas selected, as often happens within the different strata in stratified sampling. It is often left to the discretion of interviewers, etc. to include specified numbers (quotas) of subjects possessing given attributes within the sample. To do so accurately may be costly, requiring highly trained interviewers.

Quota sampling

- widely used in market research;
- sample includes a specified number or quota of subjects with given attributes;
- interviewers must be highly trained as they are often responsible for identification and selection of respondents;
- a 'biased' sample therefore results, but one which may be useful in representing the customers seen as most likely to purchase the firms products.

PAUSE FOR THOUGHT 7.2 *Can you identify three well-known products for which a quota sampling procedure might be considered appropriate for testing consumer response?*

Multi-stage and cluster sampling

Other types of non-random sampling include **multi-stage** and **cluster** sampling.

Multi-stage sampling

- usually involves sampling of subjects with a given attribute;
- often occurs where there is a wide geographical spread of such subjects which makes sampling expensive;
- whole geographical area is therefore divided into regions;
- a small number of such regions is then selected randomly;
- regions selected are further broken down into *sub-regions* (e.g. North West England into parliamentary constituencies), from which a random sample is selected;
- sub-regions selected are further broken down into *units* (e.g. towns or streets), from which a random sample is again selected;
- eventually individual households or persons with a given attribute are identified in specific towns or streets;
- costs of interviewing is then much reduced, though some risks of bias in procedure especially if only small numbers of regions, sub-regions, etc. are selected.

Cluster sampling

- items chosen in *clusters* rather than individually;
- example, cluster might be *all* residents in a particular road or group of roads;
- useful method where the population is widely spread geographically but where the various clusters are broadly representative of that population;
- similar to above sampling method, except that the cluster itself is the *single-stage* involved.

DID YOU KNOW?

If your target market were, say, boys between 11 and 14 years, a recent government report into young people's lives would provide useful marketing information. For example, boys between 11 and 14 years:

have on average £2.30 pocket money per week;
have as their favourite comics, magazines, games, and newspapers:

- *Viz (26 per cent of all boys)*
- *Gamesmaster (24 per cent of all boys)*
- *Beano (24 per cent of all boys)*
- *Sega Power (23 per cent of all boys)*
- *The Sun (22 per cent of all boys)*

have as their favourite TV programmes – Neighbours, Coronation Street, EastEnders, Home and Away (respectively).

- *have football as their main activity in free time (50 per cent of all boys) etc.*

7.2 Distribution of sample means

If the distribution of the population (X_i) is Normal then the distribution of the arithmetic means ($\bar{X}_i$) of *samples* of size n taken from that population will be Normal (see Figure 7.1). The new distribution will have the same arithmetic mean μ but a different standard deviation. We call this the *standard error* (SE) and it can be found by dividing the standard deviation of the population (σ) by the square root of the sample size (i.e. $\sqrt{n}$).

Standard error (SE) = standard deviation of distribution of sample means

$$SE = \frac{\sigma}{\sqrt{n}} \text{ or } SE = \frac{s}{\sqrt{n}}$$

Note: where the *population* standard deviation (σ) is not available, the *sample* standard deviation (s) is often used as an approximation:

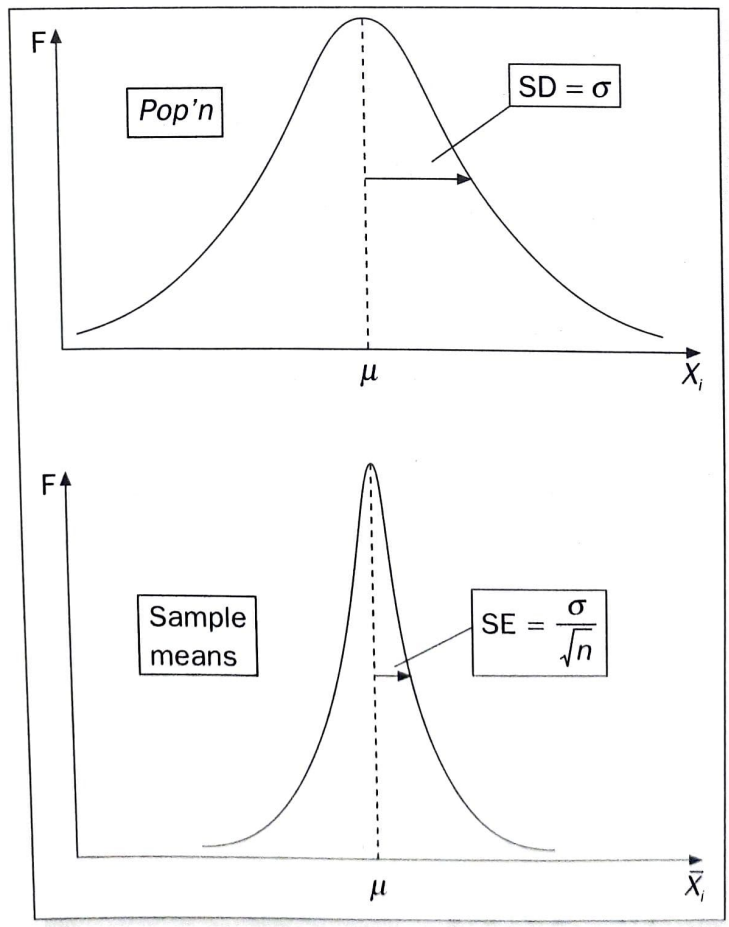


figure 7.1
Distribution of population for variable X_i and for the means $\bar{X}_i$ of samples of size n drawn from that population

Of course if the sample were, say, of size 1 then we would be back to individual members of the population X_i and there would be no difference between SD and SE (SD = SE = $\sigma/\sqrt{1} = \sigma$). However if the samples were of a size greater than 1, then the samples themselves would tend to *average out* any extreme values of X_i, so that the distribution of sample means ($\bar{X}_i$) could be expected to have a smaller dispersion around the population mean (μ) than would the distribution of each item of the population. The *larger* the sample size n, the smaller this dispersion of $\bar{X}_i$ around μ is likely to be. For example, samples of size 36 taken from the population (X_i) give an SE for the distribution of sample means ($\bar{X}_i$) of $\sigma/\sqrt{36}$ which is *smaller* than the SE had the samples taken from the population only been of size 4, giving SE = $\sigma/\sqrt{4}$.

Using the Z statistic

Because the distribution of sample means is *Normal*, we can still use the Z statistic. However note that we now calculate the Z statistic as:

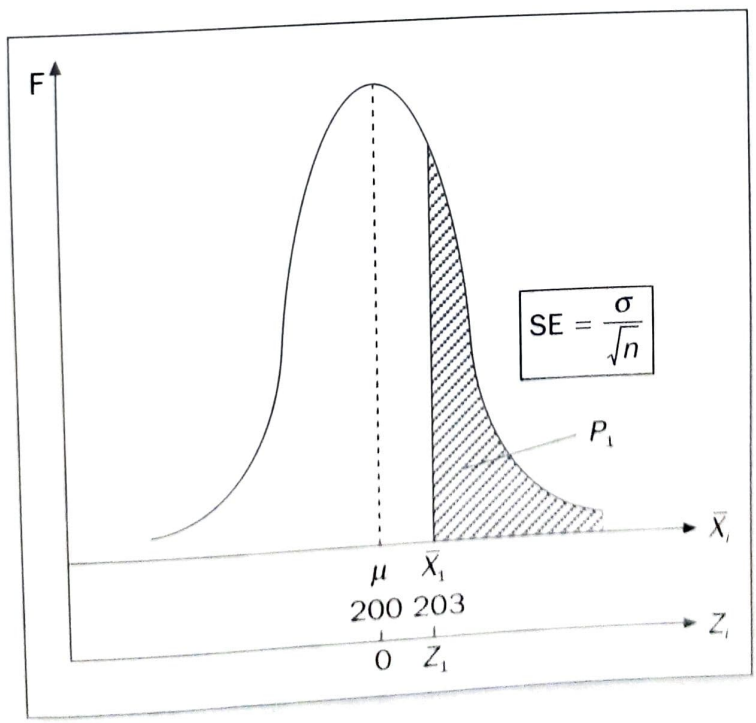
$$Z = \frac{\bar{X}_i - \mu}{\frac{\sigma}{\sqrt{n}}}$$

WORKED EXAMPLE **7.1**

A production line makes units which are normally distributed with a mean weight of 200 grams and a standard deviation (σ) of 9 grams. What is the probability of a sample of 36 units having a (sample) mean weight of 203 grams or more?

Solution

figure 7.2



μ = 200 grams

σ = 9 grams

n = sample size = 36

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{9}{\sqrt{36}} = 1.5$$

$$Z_1 = \frac{\bar{X}_1 - \mu}{\sigma/\sqrt{n}} = \frac{203 - 200}{1.5} = +2$$

Using the Z tables (Appendix 1, p. 314)

$P_1 = 0.0228$

7.3 Central Limit Theorem

Even if the population is *not* normally distributed, the **Central Limit Theorem** tells us that if sampling is random and the sample size (n) is ≥ 32, then the distribution of the sample means can be regarded as approximately normally distributed, with mean μ and standard error $\sigma/\sqrt{n}$, as shown in Figure 7.3.

We can therefore still use our Z statistic for calculating probabilities for the distribution of sample means even when the population is not itself normally distributed. All we need do is use a random sampling procedure and make sure that our sample size is large enough (at least 32).

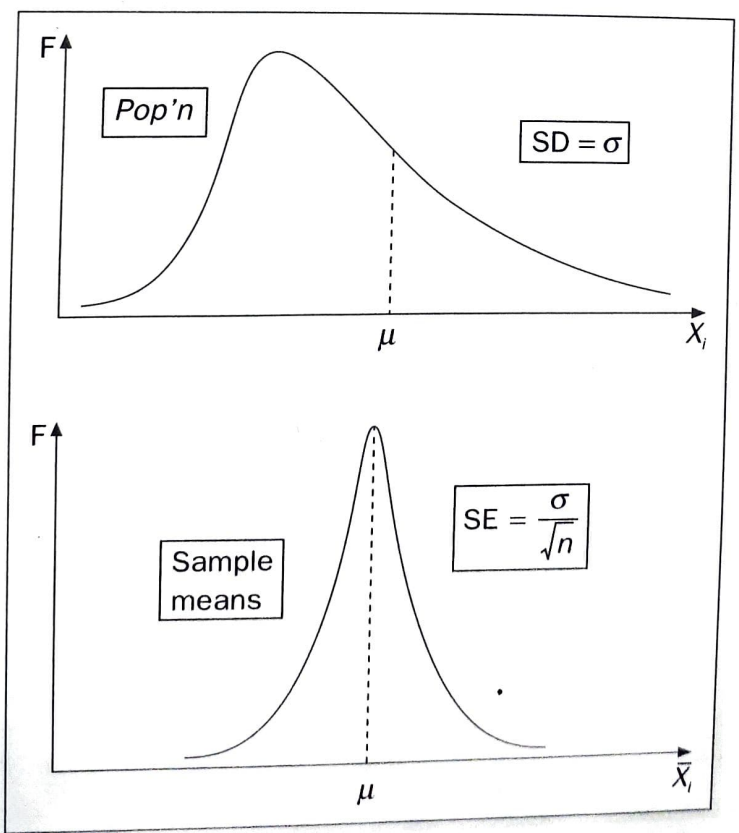


figure 7.3
Central Limit Theorem: population skewed, but distribution of sample means ($\bar{X}_i$) Normal, with same mean (μ) and SE = $\sigma/\sqrt{n}$

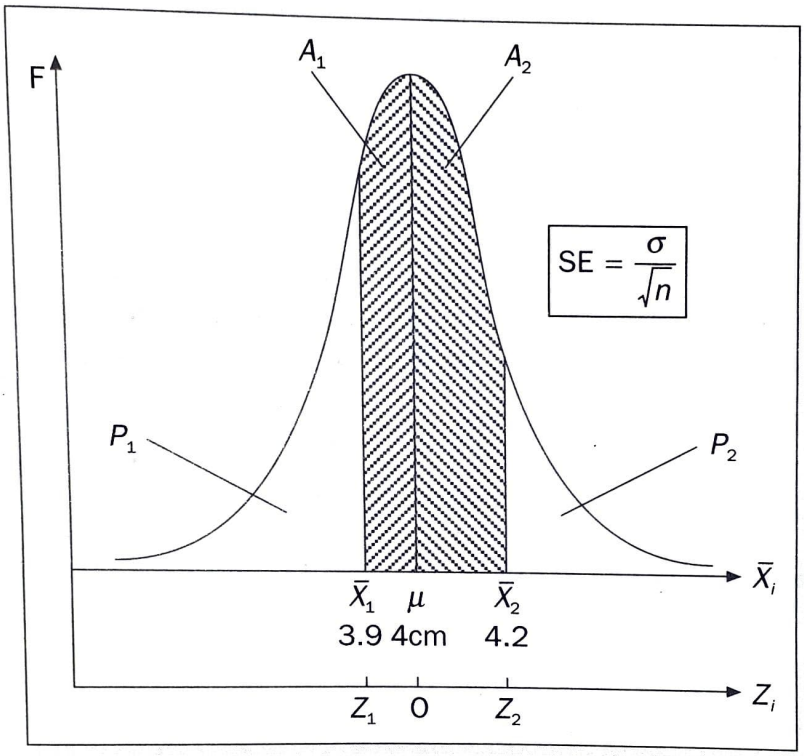
WORKED EXAMPLE **7.2**

The output of glass panels has a mean thickness of 4 cm and a standard deviation of 1 cm. If a random sample of 100 glass panels is taken, what is the probability of the sample mean having a thickness of between 3.9 cm and 4.2 cm?

Solution

The Central Limit Theorem applies as random sampling has been adopted and $n \geq 32$. The distribution of sample means ($\bar{X}_i$) is therefore Normal with mean (μ) = 4 cm and standard error $\sigma/\sqrt{n} = 1/\sqrt{100} = 0.1$ cm.

figure 7.4



$$Z_1 = \frac{\bar{X}_1 - \mu}{\sigma/\sqrt{n}} = \frac{3.9 - 4.0}{0.1} = -1$$

$P_1 = 0.1587$

$A_1 = 0.5000 - 0.1587$

$A_1 = 0.3413$

$$Z_2 = \frac{\bar{X}_2 - \mu}{\sigma/\sqrt{n}} = \frac{4.2 - 4.0}{0.1} = +2$$

$P_2 = 0.0228$

$A_2 = 0.5000 - 0.0228$

$A_2 = 0.4772$

$A_1 + A_2 = 0.8185$

i.e. there is an 81.85 per cent chance of the sample mean of 100 glass panels having a thickness of between 3.9 cm and 4.2 cm.

SELF-CHECK QUESTIONS

7.1 It is thought that the annual cost of transport to and from work for employees of a company is normally distributed with mean £100 a year and standard deviation £10. What is the probability that a *sample* of 100 employees will give a sample mean of:

(a) less than £98 a year
(b) more than £103 a year?

7.2 A departmental manager believes that the average value of a large population of invoices is £10 with a standard deviation of £2. What is the probability that a *sample* of 36 invoices selected at random will give a sample mean:

(a) below £9.50
(b) above £10.75
(c) above £11?

7.3 An automatic machine fills jars of jam with a mean net weight of 340 grams and a standard deviation of 8 grams. What is the probability that a *sample* of 64 jars selected at random will have a sample mean:

(a) above 340.75 grams
(b) below 338.5 grams
(c) between 339 grams and 342 grams?

Note: **Answers can be found on pp. 347–348.**

7.4 Confidence intervals

A **confidence interval** is a range of values within which we can have a certain level of confidence (usually 95 per cent or 99 per cent) that a particular value of a variable will lie. In the context of sampling, we are usually interested in one of two possible *types* of confidence interval.

Confidence interval for the sample mean

This is the range of values around the *population mean* (μ) within which we can be 95 per cent or 99 per cent confident that a particular *sample mean* ($\bar{X}_i$) will lie.

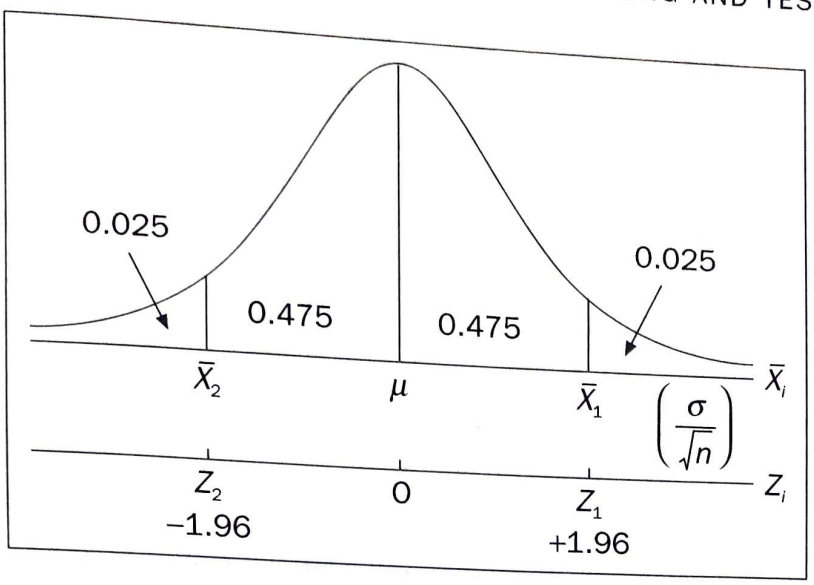
95 per cent confidence interval

In terms of Figure 7.5, we wish to find the values for $\bar{X}_1$ and $\bar{X}_2$, which are 47.5 per cent (0.475) either side of the population mean μ.

Of course this will imply 2.5 per cent (0.025) in each tail of the distribution for $\bar{X}_i$.

We have already seen (p. 130) how we calculate the Z_1 and Z_2 scores (+1.96 and −1.96 respectively) which will give us 0.025 in each tail. We can then use our formula for the Z statistic (as applied to the distribution of sample means) to calculate $\bar{X}_1$ and $\bar{X}_2$:

figure 7.5
Finding the 95% confidence interval for the sample mean ($\bar{X}_i$)



$$Z_1 = \frac{\bar{X}_1 - \mu}{\sigma/\sqrt{n}}$$

i.e. $$1.96 = \frac{\bar{X}_1 - \mu}{\sigma/\sqrt{n}}$$

$$+1.96\left(\sigma/\sqrt{n}\right) = \bar{X}_1 - \mu$$

$$\mu + 1.96\left(\sigma/\sqrt{n}\right) = \bar{X}_1$$

We can use the same procedure for finding Z_2, except that we insert −1.96 for the Z_2 score. You should then be able to show that:

$$\mu - 1.96\left(\sigma/\sqrt{n}\right) = \bar{X}_2$$

The end values of the confidence interval are called the *confidence limits*. Here $\bar{X}_1$ and $\bar{X}_2$ are the confidence limits for the 95 per cent confidence interval.

In general, the **95 per cent *confidence interval* for the sample mean** can be written as follows:

$$\bar{X}_i = \mu \pm 1.96\ (\sigma/\sqrt{n})$$

i.e. $\bar{X}_i = \mu \pm 1.96$ SE

In other words we can be 95 per cent confident that a single sample mean ($\bar{X}_i$) will lie within 1.96 standard errors (SE) of the population mean μ.

99 per cent confidence interval

The approach here is exactly the same as above *except* that we wish to find the values for the distribution of sample means, namely $\bar{X}_1$ and $\bar{X}_2$, which are 49.5 per cent (0.495) either side of the population mean μ. Of course this will imply 0.5 per cent (0.005) in each tail of the distribution for $\bar{X}_i$.

As we shall see in Figure 7.6(b) below, the only difference will be that our Z_1 and Z_2 scores respectively will be +2.58 and −2.58 in order to give 0.005 in each tail.

In general the **99 per cent *confidence interval for the sample mean*** can be written as follows:

$$\bar{X}_i = \mu \pm 2.58(\sigma/\sqrt{n})$$

i.e. $\bar{X}_i = \mu \pm 2.58$ SE

In other words we can be 99 per cent confident that a single sample mean ($\bar{X}_i$) will lie within 2.58 standard errors (SE) of the population mean (μ).

The end values of this 99 per cent confidence interval, $\bar{X}_1$ and $\bar{X}_2$, will be the 99 per cent *confidence limits*.

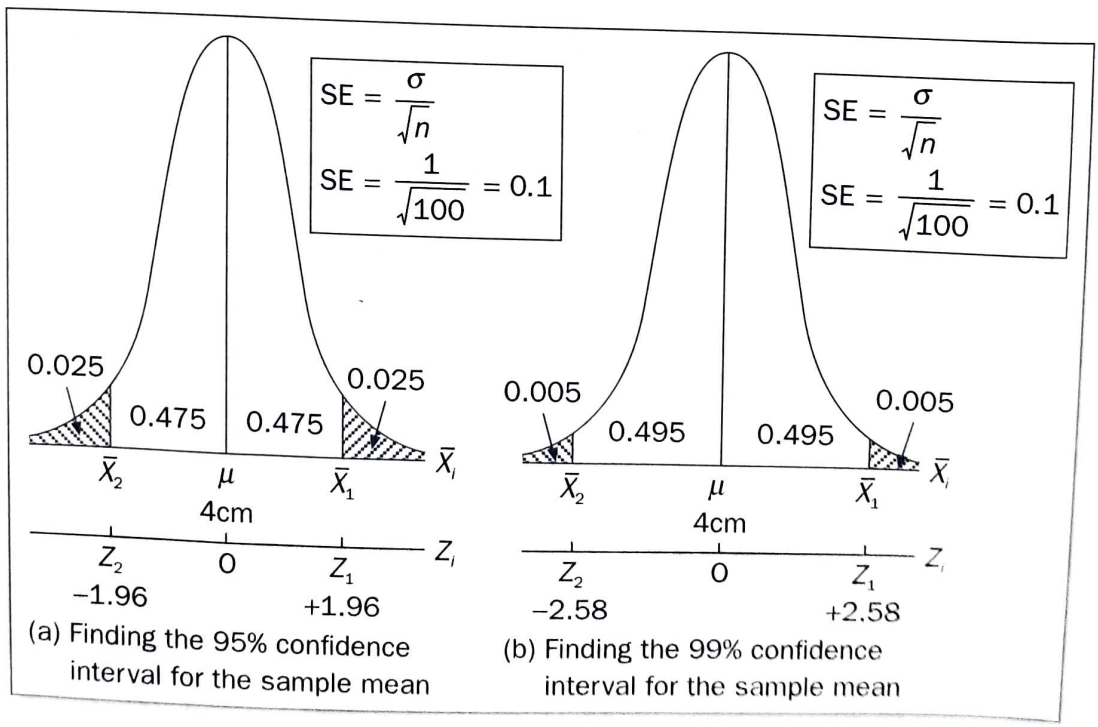
WORKED EXAMPLE **7.3**

A large number of random samples of size 100 are taken from the production line for glass panels which is thought to produce panels with mean thickness 4 cm and standard deviation 1 cm.

Find (a) the 95 per cent and (b) the 99 per cent confidence intervals and associated confidence limits for the sample mean.

Solution

figure 7.6



(a) $\bar{X}_i = \mu \pm 1.96(\sigma/\sqrt{n})$

$\bar{X}_i = 4 \pm 1.96(1/\sqrt{100})$

$\bar{X}_i = 4 \pm 1.96(0.1)$

$\bar{X}_i = 4 \pm 0.196$

$\bar{X}_i = 4.196$ and 3.804

The 95 per cent *confidence interval* is 3.804 cm to 4.196 cm

In terms of Figure 7.6(a) the *confidence limits* are:

$\bar{X}_1 = 4.196$ cm

$\bar{X}_2 = 3.804$ cm

(b) $\bar{X}_i = \mu \pm 2.58(\sigma/\sqrt{n})$

Note: if you check the Z tables in Appendix 1 (p. 314) you will see that 2.575 is strictly the Z score giving 0.005 in each tail. Rounding up the 5 in the third decimal place gives 2.58.

$\bar{X}_i = 4 \pm 2.58\ (1/\sqrt{100})$

$\bar{X}_i = 4 \pm 2.58\ (0.1)$

$\underline{\bar{X}_i = 4 \pm 0.258}$

The 99 per cent *confidence interval* is 3.742 cm to 4.258 cm.

In terms of Figure 7.6(b) the *confidence limits* are:

$\bar{X}_1 = 4.258$ cm

$\bar{X}_2 = 3.742$ cm

Confidence interval for the population mean

This is the range of values around the *sample mean* ($\bar{X}_i$) within which we can be 95 per cent or 99 per cent certain that the *population mean* (μ) will lie.

The previous analysis applies here except that in this case we are likely to be given information about a particular *sample* and asked to find the ranges of values within which we might expect to find the population mean μ.

It may be that instead of being given information about the *population standard deviation* (σ) we are only given information about the *sample standard deviation* (s). However we have already noted (p. 148) that we can use s as an approximation to σ in calculating the standard error for the distribution of sample means:

$$\text{i.e. } SE = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

We have seen that the Z statistic for the distribution of sample means is:

$$\pm Z = \frac{\bar{X}_i - \mu}{s/\sqrt{n}}$$

$$\text{i.e. } \pm Z\left(s/\sqrt{n}\right) = \bar{X}_i - \mu$$

$$\mu = \bar{X}_i \pm Z\left(\frac{s}{\sqrt{n}}\right)$$

All we have done is re-arrange the terms to have μ on the left-hand side this time instead of $\bar{X}_i$. As before the **confidence intervals for population mean** can be written as follows:

$$\mu = \bar{X}_i \pm 1.96\left(\frac{s}{\sqrt{n}}\right) \text{ for a 95\% confidence interval}$$

$$\mu = \bar{X}_i \pm 2.58\left(\frac{s}{\sqrt{n}}\right) \text{ for a 99\% confidence interval}$$

WORKED EXAMPLE **7.4**

Estimate (a) the 95 per cent and (b) the 99 per cent confidence intervals and associated confidence limits for the *population mean* when we are given the following *sample data*:

sample mean ($\bar{X}_i$) = 950 kg
sample standard deviation (s) = 15 kg
sample size (n) = 36

Solution

(a) *95 per cent confidence interval and confidence limits for μ*

$\mu = \bar{X}_i \pm 1.96(s/\sqrt{n})$

$\mu = 950 \pm 1.96(15/\sqrt{36})$

$\mu = 950 \pm 1.96(2.5)$

$\mu = 950 \pm 4.9$

μ = 945.1 kg to 954.9 kg

In other words we can be 95 per cent confident that the population mean μ will lie in the interval between 945.1 kg and 954.9 kg, given the sample data provided. These values are the 95 per cent *confidence limits* for the population mean.

(b) *99 per cent confidence interval and confidence limits for μ*

$\mu = \bar{X}_i \pm 2.58(s/\sqrt{n})$

$\mu = 950 \pm 2.58(15/\sqrt{36})$

$\mu = 950 \pm 2.58(2.5)$

$\mu = 950 \pm 6.45$

μ = 943.55 kg to 956.45 kg

We can be 99 per cent confident that the population mean μ will lie in the interval between 943.55 kg and 956.45 kg, given the sample data. These values are the 99 per cent *confidence limits* for the population mean.

SELF-CHECK QUESTIONS

7.4 The management of a company claim that the average weekly earnings of their employees is £550 with a standard deviation of £120. A random sample of 144 employees is taken. Calculate (a) the 95 per cent and (b) the 99 per cent confidence intervals for the *sample* mean.

7.5 A random sample of 64 households was selected in a town and the average expenditure per week on a certain product was found to be £6.15 with a (sample) standard deviation of £1.30. Calculate (a) the 95 per cent and (b) the 99 per cent confidence intervals for the *population* mean.

Note: **Answers can be found on p. 349.**

ACTIVITY 5

Confidence interval for μ when s is unknown

Two random samples of components were selected from a production process with the intention of estimating their life expectancy in hours. The first random sample contained 60 components, the second random sample contained 120 components. Estimate the 90 per cent, 95 per cent and 99 per cent confidence intervals for the population mean from each sample.

The data from the samples were as follows:

Sample 1

286.5	304.5	301.3	313.9	309.7	287.4
310	310	306.4	288.4	310.4	285.7
300.6	296.5	304.6	291.5	299.2	296.7
314.4	307.2	299.8	316.1	287	307.9
293.7	302.4	312.4	302.1	296.4	292.7
287.5	310.1	305.5	295.6	291	301.2
298.5	299.7	307.7	304.3	292.6	296.2
304.6	302	290.8	314	298.5	291.6
301.9	297.2	298.2	302.1	295.2	299.7
292.9	294.6	296.2	306.5	284.2	304.6

Sample 2

297	311.9	301.4	305	305.1	309.7	295.1	319.4
310.5	301.2	313.4	288.4	301.8	295.2	305.8	289.2
309.8	295	301.3	303.7	304.3	290.2	298.3	304.7
298.6	294.9	292.2	301.6	309.2	305.7	312.8	301.2
291.9	305.8	290	300	298.5	302.1	301.8	299.1
295.2	321.6	297.9	303.9	293.9	300.6	301.5	311.3
291.5	301.2	310.4	299.4	308.9	299.6	291.1	287.7
286.6	290.1	305.7	306.2	304.8	301.6	299.3	307.9
297.4	297.4	317.2	302.4	305.5	290	313.6	292.1
309.6	295.4	309.8	299.4	289.5	307.8	299.1	296.1
291.4	298.1	292.7	292.6	277	301.8	291.5	308.6
297.7	302.3	293.2	289	300.1	303.8	297.2	276.4
304.8	296.7	297.4	290.2	299.7	303.1	288.6	293.9
304.4	304.8	303.2	299.1	290.8	304	286	308.2
306.2	304.4	297.2	296.6	315.7	302.6	297.9	288.3

1 Enter all the data from Sample 1 into column **A** and the data from Sample 2 into column **B**.

2 Analyse the two samples by setting out a table of results as follows:

	D	*E*	*F*
2	**SAMPLE**	**1**	**2**
3	Size		
4	Median		
5	Mean		
6	Std Dev		
7	Std Error		
8	Variance		
9	Min		
10	Max		
11	Range		

3 Comment on your table of results.

4 Calculate the standard error by using: $SE = \frac{s}{\sqrt{n}}$

5 Applying the Central Limit Theorem we can say that for 95 per cent:

$$\mu = \bar{X} \pm 1.96\frac{s}{\sqrt{n}}$$

For 90 per cent and 99 per cent:

$\mu = \bar{X} \pm 1.645\frac{s}{\sqrt{n}}$ and $\mu = \bar{X} \pm 2.575\frac{s}{\sqrt{n}}$ respectively.

6 Use the above formulæ to find the respective confidence intervals. It is suggested that another table, such as the one below, will produce a clea. layout for your solutions.

Conf level %	*Z*	*Z*E*	*Lower*	*Upper*	*Interval*
Sample 1					
90	1.645				
95	1.96				
99	2.575				
Sample 2					
90	1.645				
95	1.96				
99	2.575				

Comment on your results. Check with the responses on p. 468.

7.5 Tests of hypotheses: principles

We can use our work on the distribution of sample means and the Central Limit Theorem for testing **hypotheses** (theories). A hypothesis is an assumption about a situation. We usually want to *test* this assumption against one or more alternative assumptions.

It will be useful at this stage to become familiar with the terminology widely used in this topic area. For simplicity we assume that we are testing a main hypothesis (*Null* hypothesis, H_0) against one other hypothesis (*Alternative* hypothesis, H_1).

Null hypothesis (H_0): the hypothesis to be tested.

For example that the mean thickness of glass panels is 4 cm:

$H_0 : \mu = 4$ cm

Alternative hypothesis (H_1): the hypothesis we wish to accept if we reject the null hypothesis.

For example that the mean thickness of glass panels is *not* equal to 4 cm.

$H_1 : \mu \neq 4$ cm

In testing our null hypothesis, there are essentially two *types of error* we might wish to avoid.

- **Type 1 error:** rejecting the null hypothesis when it is, in fact, true.
- **Type 2 error:** accepting the null hypothesis when it is, in fact, false.

Avoiding a Type 1 error is usually the main concern of problems in this topic area. When testing the null hypothesis we usually state the *maximum risk* we are willing to take of committing a Type 1 error, i.e. of rejecting the null hypothesis when it is, in fact, true. This is called the *level of significance* and is typically either 5 per cent (0.05) or 1 per cent (0.01).

Levels of significance

The maximum risk we are willing to take of making a Type 1 error, i.e. of rejecting the null hypothesis when it is, in fact, true.

Confidence limits and critical values

If we look back to Figure 7.6 (p. 154) we can usefully illustrate the idea of the level of significance. It is clear that there is a 95 per cent (0.95) chance of any sample mean ($\bar{X}_i$) lying within 1.96 standard errors (SE) of the true population mean μ. However there is still a 5 per cent (0.05) chance of a single sample mean lying *outside* this 95 per cent *confidence interval* even though the population mean really is μ.

Here this 5 per cent (0.05) chance is split evenly to be 2.5 per cent (0.025) in each tail. If the sample mean lies outside the *confidence limits* $\bar{X}_1$ or $\bar{X}_2$ then our decision will be to reject H_0 even though we might be wrong in rejecting H_0. However the chance of our being wrong is *less than 5 per cent*, since there was a less than 5 per cent chance of actually getting a sample mean $\bar{X}_i$ outside $\bar{X}_1$ and $\bar{X}_2$ when μ really is the population mean.

These *confidence limits* for the population mean, μ, can therefore be regarded as the *critical values* for tests of hypotheses. In other words they are the values of $\bar{X}_i$ (or their Z_i equivalents) *outside of which* we are willing to reject the null hypothesis (H_0). In that sense they are the values 'critical' to the decision we are about to take in our test of the null hypothesis (H_0).

Critical values

These are the values of X_i or Z_i for the sample *outside* of which we are willing to reject the null hypothesis (H_0) at a particular level of significance.

Two- and one-tailed tests

Two-tailed tests

The approach so far has involved a **two-tailed test**, i.e. where we split the risk of being wrong in rejecting the null hypothesis *equally* between each tail of the distribution.

Let us formally set out the null (H_0) and alternative (H_1) hypotheses for a *two-tailed* test. We use our earlier example of thickness of glass panels by way of illustration.

$H_0 : \mu = 4$ cm

$H_1 : \mu \neq 4$ cm

Since the alternative hypothesis (H_1) says $\mu \neq 4$ cm it is clear that we are just as interested in situations where $\mu > 4$ cm as we are in situations where $\mu < 4$ cm.

Two-tailed test

We divide the risk we are willing to take in making a Type 1 error equally between each tail of the distribution.

If we test the null hypothesis at a 5 per cent (0.05) level of significance, we place 2.5 per cent (0.025) in each tail. Here we will reject H_0 if we think that $\mu > 4$ cm or if we think that $\mu < 4$ cm.

Acceptance and rejection regions

- The region *within which* we are willing to accept the null hypothesis (H_0) we call the *acceptance region*.
- The region *outside which* we are willing to reject the null hypothesis (H_0) we call the *rejection region*.

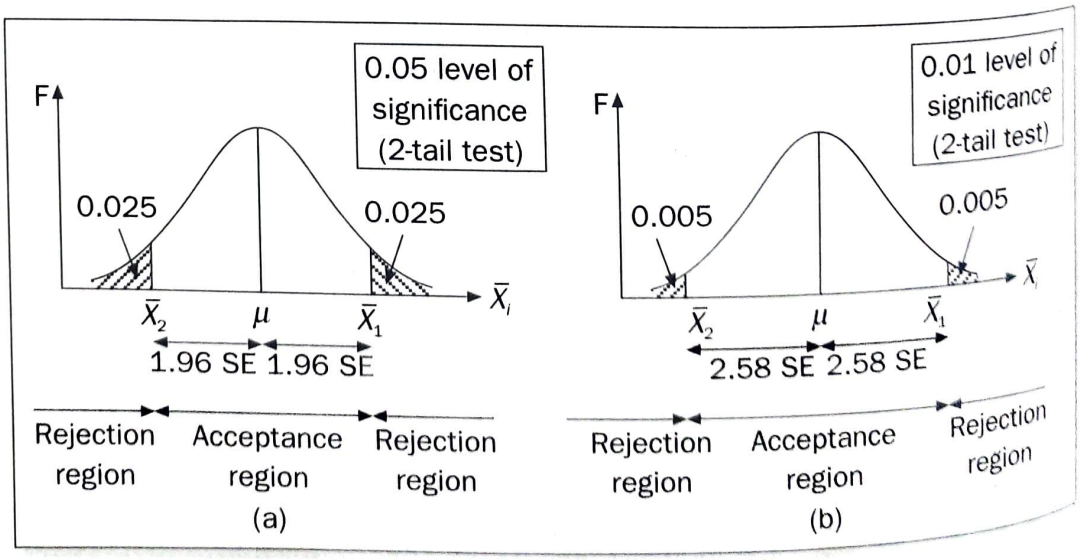


figure 7.7
Acceptance and rejection regions for the null hypothesis (H_0): two tailed tests at (a) 5% (0.05) (b) 1% (0.01) levels of significance

$\bar{X}_1$ and $\bar{X}_2$ correspond to the **critical values**, i.e. the sample means at the boundary between the acceptance and rejection regions for the null hypothesis (H_0).

WORKED EXAMPLE **7.5**

Find the critical values for a two-tailed test at the 5 per cent (0.05) level of significance for the earlier example of glass panels, mean (μ) = 4 cm, standard error ($\sigma/\sqrt{n}$) = 0.1 cm.

Solution

$\mu \pm 1.96$ SE = critical values

$4 \pm 1.96(0.1)$ = critical values

3.804 cm = $\bar{X}_2$; 4.196 cm = $\bar{X}_1$

We can be 95 per cent certain of getting a sample mean in the *acceptance region* 3.804 cm to 4.196 cm if the true (population) mean is 4 cm.

If, however, we get a sample measurement *outside* these critical values (i.e. in the *rejection region*), then we can reject $H_0 : \mu = 4$ cm with less than a 5 per cent (0.05) chance of being wrong. This is because there is a less than 5 per cent (0.05) chance of $\bar{X}_i$ being outside these critical values if $H_0 : \mu = 4$ cm is true.

WORKED EXAMPLE **7.6**

Find the critical values for a two-tailed test at the 1 per cent (0.01) level of significance for the earlier example of glass panels, mean (μ) = 4 cm, standard error ($\sigma/\sqrt{n}$) = 0.1 cm.

Solution

This time we place half of 1 per cent (0.005) in each tail, giving a Z score of ±2.58.

$\mu \pm 2.58$ SE = critical values

$4 \pm 2.58(0.1)$ = critical values

3.742 cm = $\bar{X}_2$; 4.258 cm = $\bar{X}_1$

We can be 99 per cent certain of getting a sample mean in the *acceptance region* 3.742 cm to 4.258 cm if the true (population) mean is 4 cm.

If however, we get a sample measurement *outside* these critical values (i.e. in the *rejection region*), then we can reject $H_0 : \mu = 4$ cm with less than a 1 per cent (0.01) chance of being wrong. This is because there is a less than 1 per cent chance of $\bar{X}_i$ being outside these critical values if $H_0 : \mu = 4$ cm is true.

One-tailed test

Here we do *not* split the risk of being wrong in rejecting the null hypothesis equally between each tail of the distribution. Here we are only concerned

with *one side* (and therefore *one tail*) of the distribution when testing our null hypothesis (H_0) against an alternative hypothesis (H_1). We therefore place *all* the risk of making a Type 1 error in one tail.

One-tailed test

We place the risk we are willing to take in making a Type 1 error in one tail of the distribution.

Let us formally set out the null (H_0) and alternative (H_1) hypotheses for a **one-tailed** test. We again use our earlier example of thickness of glass panels by way of illustration.

EITHER

$H_0 : \mu = 4$ cm

$H_1 : \mu > 4$ cm

In testing H_0 against H_1 we are only interested in rejecting H_0 (i.e. accepting H_1) if the sample means suggest a true (population) mean *greater than* 4 cm.

OR

$H_0 : \mu = 4$ cm

$H_1 : \mu < 4$ cm

In testing H_0 against H_1 we are only interested in rejecting H_0 (i.e. accepting H_1) if the sample mean suggests a true (population) mean *less than* 4 cm.

Acceptance and rejection regions

Figure 7.8 indicates the relevant acceptance and rejection regions for a one-tail test at 5 per cent (0.05) and 1 per cent (0.01) levels of significance. Remember the Z scores are calculated using the Z tables (Appendix 1, p. 314) and are used to give the area in the respective tails.

WORKED EXAMPLE **7.7**

Look again at Figure 7.8(a), p. 163. Find the critical values for a one-tailed test at the 5 per cent (0.05) level of significance for the earlier example of glass panels, mean (μ) = 4 cm, standard error ($\sigma/\sqrt{n}$) = 0.1 cm.

Solution

$\bar{X}_1 = \mu + 1.65$ SE

$\bar{X}_1 = 4 + 1.65(0.1)$

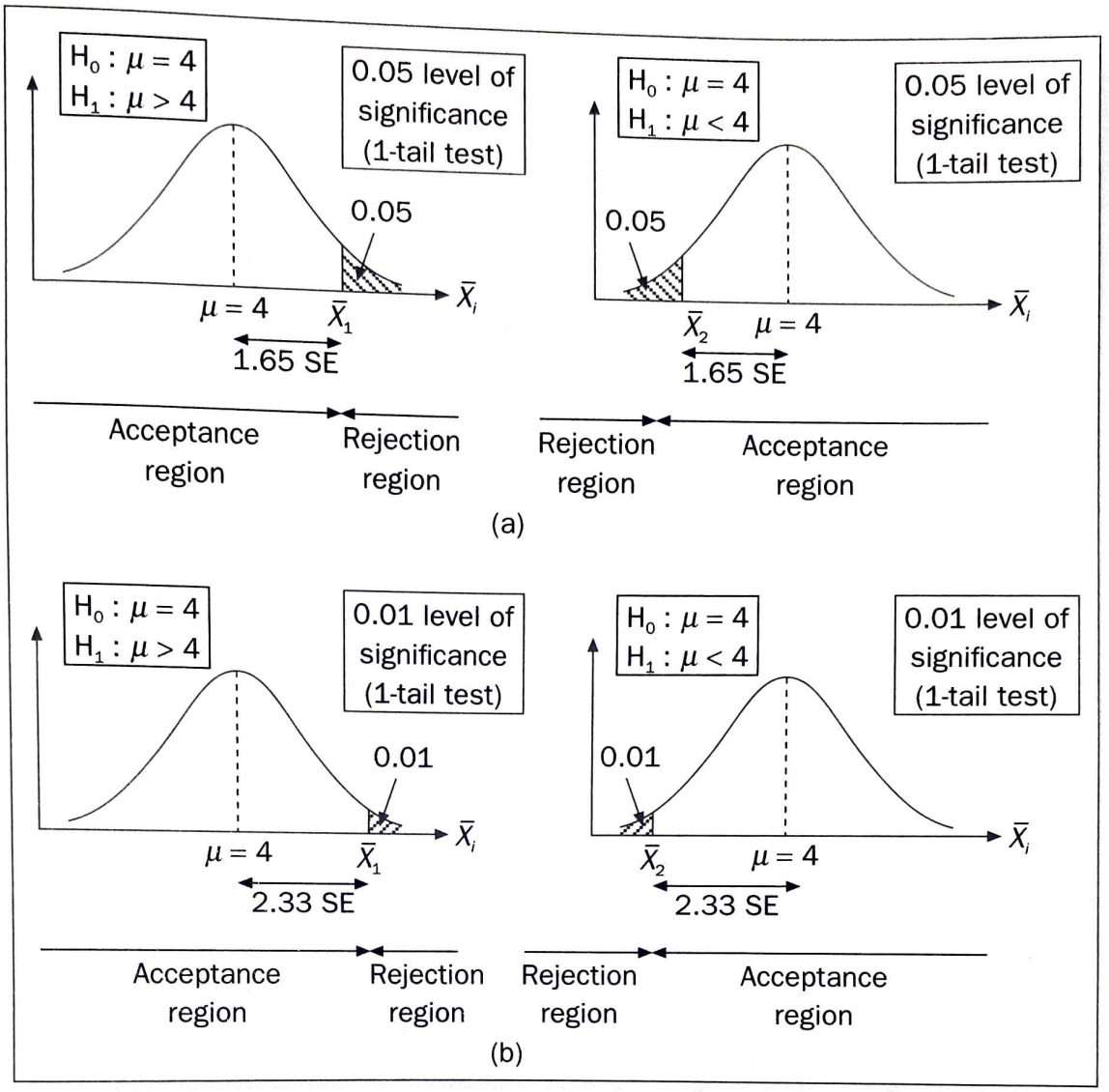
$\bar{X}_1 = 4.165$ cm

$\bar{X}_2 = \mu - 1.65$ SE

$\bar{X}_2 = 4 - 1.65(0.1)$

$\bar{X}_2 = 3.835$ cm

figure 7.8
Acceptance and rejection regions for the null hypotheses (H_0): one-tailed tests at (a) 5% (0.05) (b) 1% (0.01) levels of significance



If the sample mean $\bar{X}_i$ lies *outside* these critical values then we would be outside the *acceptance region* and in the *rejection region*.

▶ For $H_0 : \mu = 4$ cm
$H_1 : \mu > 4$ cm

Reject H_0 if $\bar{X}_i > 4.165$ cm

▶ For $H_0 : \mu = 4$ cm
$H_1 : \mu < 4$ cm

Reject H_0 if $\bar{X}_i < 3.835$ cm

7.6 Tests of hypotheses: practice

In this section we develop a simple seven-step plan for tackling all problems involving tests of hypotheses. This seven-step plan draws on the ideas presented in the previous section.

However before we outline these steps let us select *one* of two alternative approaches to finding critical values.

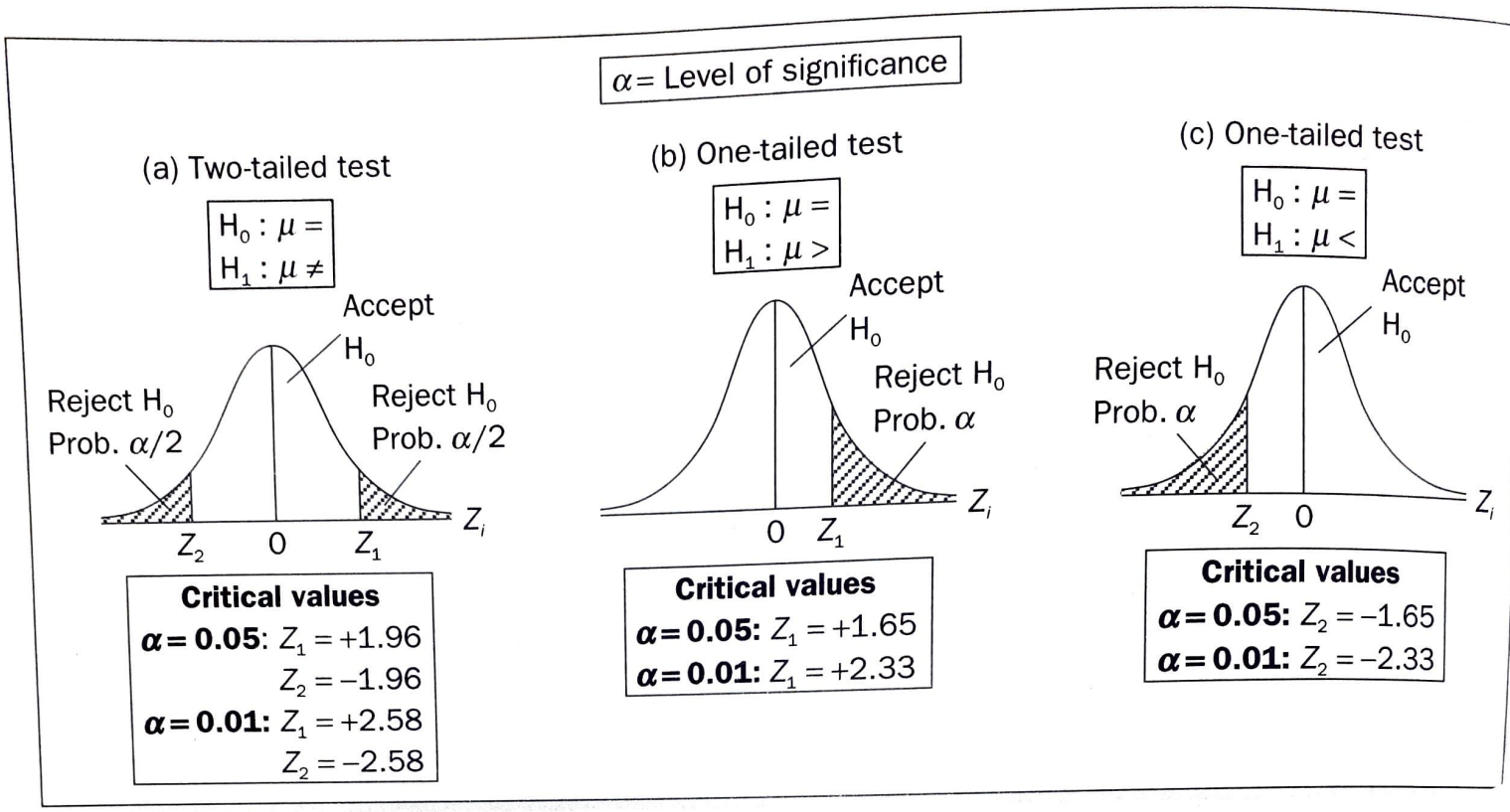


figure 7.9
Using the Z scores for the critical values
(a) Two-tailed test
(b) One-tailed test
(c) One-tailed test

Critical values: Z scores or sample mean values

As we saw in Figure 7.8 (p. 163), the critical values can be expressed in terms of the *sample means*, $\bar{X}_1$ and $\bar{X}_2$ respectively. However they can just as easily be expressed in terms of the *Z scores*, i.e. the number of standard errors from μ, the population mean.

In Figure 7.9, α (alpha) is the *level of significance* at which the test is conducted. The *Z scores* are shown for the respective *critical values*. These Z scores can be said to depend on both the level of significance and on whether the test is one- or two-tailed.

In this approach we calculate the *Z-score* using the *sample data* provided. If the Z score for the sample data lies *outside* the relevant Z critical values, then the decision rule will be to *reject* the null hypothesis (H_0).

It will be useful at this stage to go through a worked example before using this to illustrate our seven-step plan for testing hypotheses.

WORKED EXAMPLE **7.8**

It is thought that the average wage for a large company is £130 a week with a standard deviation of £30. A sample of 100 employees finds the sample average to be £123.

Solution

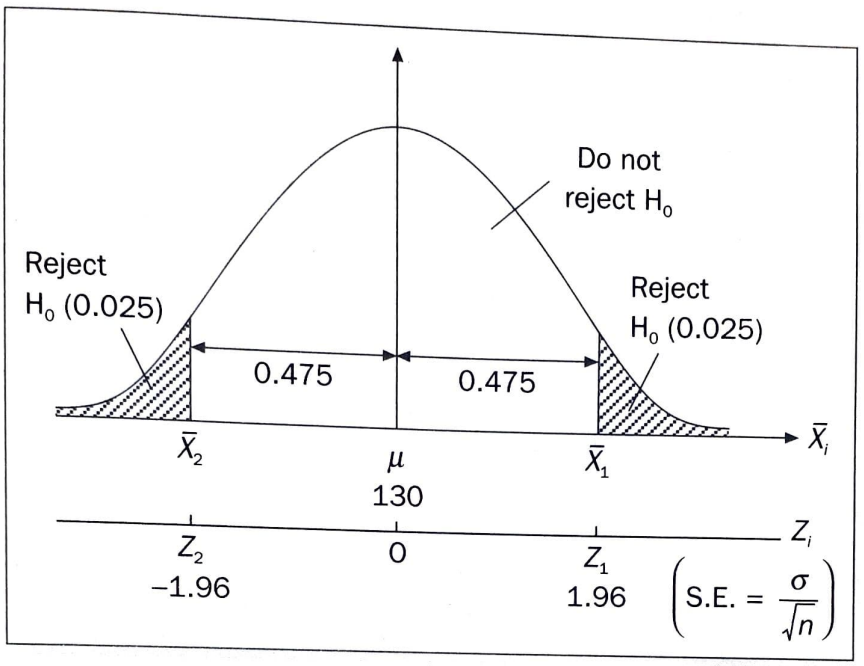
Test $H_0 : \mu = £130$
$H_1 : \mu \neq £130$

using a 5 per cent level of significance.

This is a two-tailed test since the alternative hypothesis (H_1) has been expressed as $\neq$.

If the Z score is calculated for the sample data and found to be outside the Z *critical value*, then reject H_0.

figure 7.10



Here $\bar{X} = 123$,

gives $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

$$Z = \frac{123 - 130}{30/\sqrt{100}}$$

i.e. $$Z = \frac{123 - 130}{3}$$

$$Z = -\frac{7}{3}$$

$$\underline{Z = -2.33}$$

Our Z score is *outside* the Z critical value of −1.96, so we reject H_0.

We can now formally set out our *seven-step plan* for testing a hypothesis, using this example to illustrate each step.

Step		Example
1	State hypotheses	$H_0 : \mu = £130$ $H_1 : \mu \neq £130$
2	State significance level	5% (0.05)
3	State critical values	$Z_1 = +1.96$ $Z_2 = -1.96$
4	Calculate the Z score for the *sample*	$Z = -2.33$
5	Compare this Z sample score with the Z critical values	Here $Z = -2.33$ is *outside* $Z_2 = -1.96$
6	Come to a conclusion	Reject H_0
7	Put your conclusion in words	The sample evidence does *not* support the null hypothesis at a 5 per cent level of significance

Note: A sketch diagram will usually help in Step 3 (e.g. Figure 7.9(a) for $\alpha = 0.05$)

It will be useful for you to see this seven-step plan applied to a couple of examples, before trying it out yourself.

WORKED EXAMPLE **7.9**

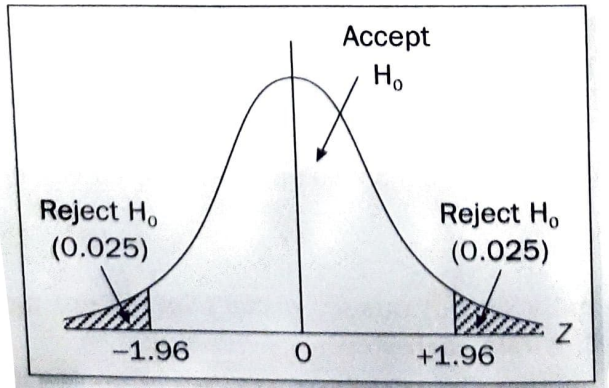
A manager of a health store believes the average amount spent per week on vitamin C supplements to be £2.50. A random sample of 100 shoppers found an average expenditure on such supplements of £2.40 with a standard deviation of £0.40. Test the manager's belief at a 5 per cent level of significance.

Solution

Note: We shall use a *two-tailed* test here, since any divergence from £2.50 (above or below) would invalidate the manager's belief. Since only the sample mean (s) is available, we use this instead of the population mean (σ) in calculating the standard error.

Step 1 $H_0 : \mu = £2.50$
$H_1 : \mu \neq £2.50$
Step 2 5 per cent (0.05) level of significance
Step 3 Critical values for Z: ±1.96

figure 7.11



Step 4 For sample data

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{2.40 - 2.50}{0.40/\sqrt{100}}$$

$$Z = \frac{-0.10}{0.40/10} = \frac{-0.10}{0.04} = -2.50$$

Step 5 −2.50 is *outside* −1.96
Step 6 Reject null hypothesis (H_0)
Step 7 Evidence does not support the health store manager's belief that average spending on Vitamin C supplement is £2.50 per week.

WORKED EXAMPLE **7.10**

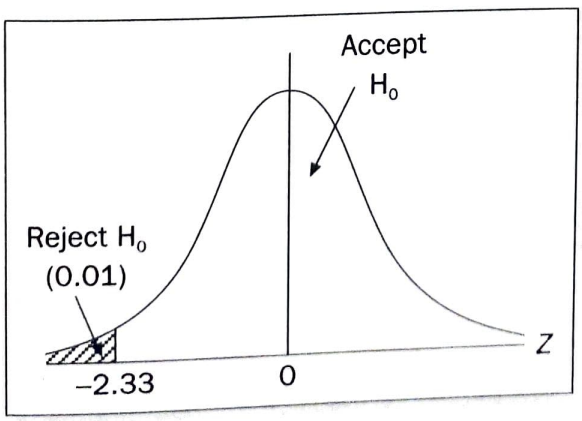
The expected mean diameter of a batch of cables is 2 cm. A quality control inspector takes a sample of 64 cables and finds that the sample has a mean diameter of 1.94 cm and a standard deviation of 0.4 cm. Test whether the manufacturing process is now producing cables with a smaller diameter than previously at a 1 per cent level of significance.

Solution

Note: We shall use a *one-tailed* test here, since our concern is that the manufacturing process is producing *smaller* diameter cables than claimed. Again we use sample standard deviation (s) as an approximation for the population standard deviation (σ). Our test is at the 1 per cent level of significance this time; i.e. we are less willing than in the previous worked example to take the risk of rejecting the null hypothesis (H_0) when it is in fact true.

Step 1 $H_0 : \mu = 2.00$ cm
$H_1 : \mu < 2.00$ cm
Step 2 1 per cent (0.01) level of significance
Step 3 Critical value for Z: −2.33

figure 7.12



Step 4 For sample data

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{1.94 - 2.00}{0.4/\sqrt{64}}$$

$$Z = \frac{-0.06}{0.4/8} = \frac{-0.06}{0.05} = -1.2$$

Step 5 –1.2 is *inside* –2.33

Step 6 Accept null hypothesis (H_0)

Step 7 There is insufficient evidence from the sample data to reject the initial hypothesis that the manufacturing process is still producing cables of diameter of 2 cm. We would be taking more risk than we are willing to take if we were to accept the alternative hypothesis that the cables are now being produced with a diameter of less than 2 cm.

You can now test your familiarity with this *seven-step* approach with the following Self-check questions. There are also more questions for practice (Review questions) at the end of this chapter.

SELF-CHECK QUESTIONS

7.6 A firm manufactures light bulbs and the sales manager claims that tests have proved that the average life of these bulbs is 1,000 hours. We wish to buy some of these light bulbs but before doing so we wish to check the claim that their average life is correct. We therefore take a sample of 36 bulbs and test them; we find that the average life of the bulbs in the sample is 940 hours with a standard deviation of 126 hours. Is the manufacture's claim justified? Test at both (a) 5 per cent (0.05) and (b) 1 per cent (0.01) levels of significance.

7.7 A factory manager asserts that the mean running time of machines is 14 hours a day. A random sample of 64 machines shows that their mean running time was only 13 hours 20 minutes, with a standard deviation of 3 hours. Test the manager's assertion at the 5 per cent significance level.

***Note:* Keep the units the same throughout (i.e. hours or minutes).**

7.8 A machine produces electronic components whose lifetimes have a mean of 100 hours and a standard deviation of 14 hours. A new technique is now introduced into their production. A *sample* of 49 components gives a mean of 104 hours. Has the new technique produced longer lasting components? Test at a 5 per cent level of significance.

***Note:* Answers can be found on pp. 349–351.**

ACTIVITY 6

Hypothesis testing or significance testing

A company believes that the average amount spent by customers has changed from its original value. The accounting department, however, disputes this claim and maintains that sales remain at £1.48 per customer. To support their view the accounting department quote the following random sample of spending (£) by 75 customers.

1.61	1.16	1.72	1.11	1.58	1.44	1.37	1.60	1.71	1.10
1.50	1.87	1.54	1.09	1.41	1.52	1.59	1.10	1.40	1.60
1.68	1.30	1.69	1.55	1.59	1.54	1.80	1.55	1.52	1.37
1.15	1.44	1.43	1.81	1.53	1.03	1.61	1.72	1.29	1.35
1.09	0.88	0.90	1.06	1.83	1.80	1.33	1.91	1.31	1.42
1.86	1.62	1.71	0.95	1.58	1.45	1.59	1.79	1.01	1.13
1.53	1.46	1.67	1.67	1.53	1.75	0.94	1.01	1.42	2.07
1.23	1.23	1.41	1.46	1.19					

Use a spreadsheet with which you are familiar to conduct the following tests.

1 Test, at the 5 per cent significance level, the hypothesis that the average amount spent by each customer remains at £1.48.
2 Test the same hypothesis at the 1 per cent significance level.

Conduct the hypothesis test:

Step 1 State the hypothesis: $H_0 : \mu = £1.48$: $H_1 : \mu \neq £1.48$

	F	G	H	I
2	**H_0:**	Average equals		1.48
3	**H_1:**	Average is not equal		1.48

Step 2 State the significance level: 0.01, 0.05 or 0.1 (1%, 5% or 10%) – for this example use 0.05

	F	G
5	**Sig level**	0.05

Step 3 State the critical values

	F	G	H
7	**Critical values**	**Lower**	**Upper**
8		=NORMSINV($G$5/2)	=1–NORMSINV($G$5/2)

Step 4 Calculate the Z statistic using the formula:

$$Z = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{\textit{sample mean} - \textit{hypothesised mean}}{\textit{standard error}}$$

	F	G
10	**Z statistic**	=(D4–I2)/D6

Step 5 Compare the Z value to the critical values.
Step 6 Come to a conclusion. If the Z value falls between the critical values then ***accept*** the null hypothesis that $\mu = £1.48$.

The decision on whether to accept or reject the null hypothesis can also be made by the computer using the logical operators **IF()**, and **OR()**.

The **IF** function returns one value if the logical_value evaluates to **TRUE** and another value if it evaluates to **FALSE** i.e.:

=IF(logical_test,*value_if_true*,*value_if_false*) or
=IF(True, *Then return this, Else return this*)

The **OR** function returns **TRUE** if one or more of the arguments is true and returns **FALSE** if all the arguments are false.

=OR(logical 1,logical 2**)**

These two functions can be combined to make the decision – thus:

	F	*G*
12	**Compare**	=IF(OR(G10<G8, G10>H8), 'REJECT','ACCEPT')

This can be translated as:
If the value in cell G10 is less than that of G8 OR greater than that of cell H8 then return REJECT else return ACCEPT.

Now try changing the value in cell **G5** to 0.1.

This construction can only be used for two-tailed tests although combinations of the logical operators can be used to develop both one-tailed tests and two-tailed tests. A response to this activity can be found on p. 469.

7.7 Student *t* distribution

For repeated *small* samples ($n < 32$) it is no longer accurate to assume that the distribution of sample means follows a *Normal* distribution. As a result we can no longer use the Z tables to work out the probabilities. Instead, we assume that the distribution of sample means follows the *Student t* distribution (called '*t*' distribution for simplicity). This distribution actually changes as the sample size changes. We must therefore apply different probabilities to samples taken with different sizes. The concept of *degrees of freedom* (v) captures this idea, where the number of degrees of freedom is expressed as $v = n - 1$, where n is sample size (see Figure 7.13).

We can illustrate this idea using 2½ per cent (0.025) in the tail of the t distribution:

$v\ (n-1)$	0.025
1	12.706
2	4.303
5	2.571
10	2.228
30	2.042

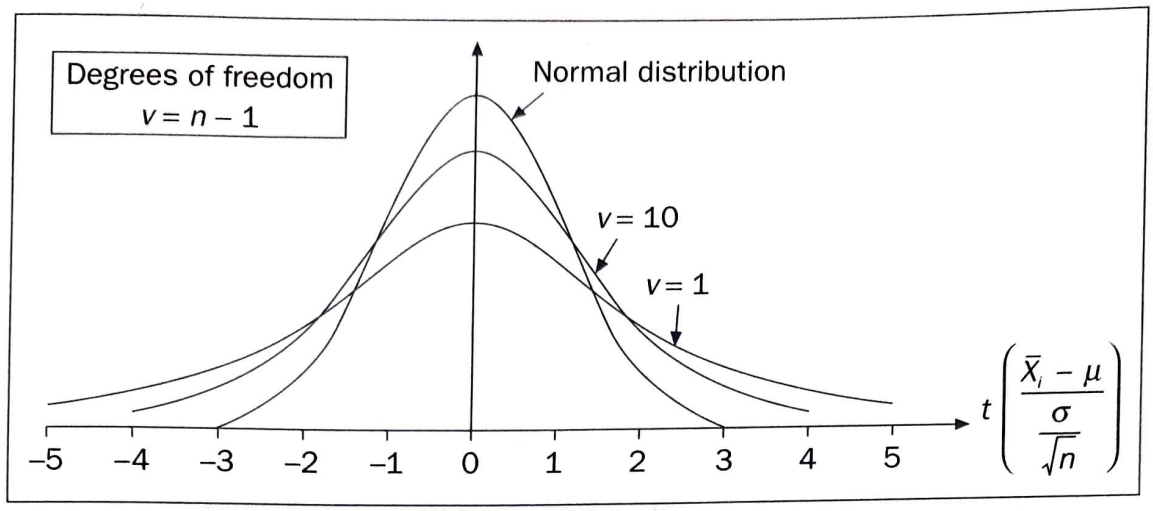


figure 7.13
Showing how the distribution of sample means ($\bar{X}_i$) expressed in terms of standard errors (t), takes different shapes depending on the number of degrees of freedom v (= n − 1)

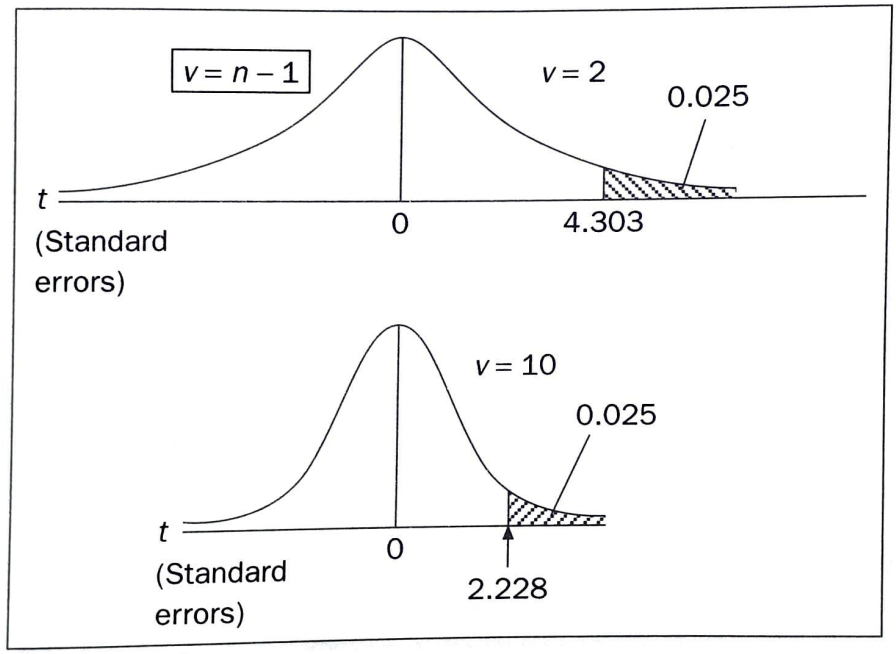


figure 7.14
Showing how the t score for 2½% (0.025) in the right-hand tail will vary depending on the number of degrees of freedom v (= n − 1)

Notice that the *critical values* (i.e. number of *standard errors* from the mean) vary depending on the value of v (= n − 1). So with a sample size (n) of 3 and v (= n − 1) of 2, we must be 4.303 standard error from the mean to get 2½ per cent (0.025) in either tail. However with a sample size (n) of 11 and v (= n − 1) of 10, we need only be 2.228 standard errors from the mean to get 2½ per cent (0.025) in either tail.

Critical values

We express the **critical values** for the t distribution as

$t_{\alpha,v}$

where α = level of significance shown in the tables
v = number of degrees of freedom

So for $\alpha = 0.025$, and $v = 2$ in the table above, we can write.

$t_{\alpha,\nu} = t_{0.025,2} = 4.303$

and when $\nu = 10$

$t_{\alpha,\nu} = t_{0.025,10} = 2.228$

t statistic

We calculate the ***t* statistic** or ***t* score** exactly as we did for Z, i.e. we use the *sample* data to find the number of *standard errors* from the mean:

$$t = \frac{\bar{X}_i - \mu}{\sigma/\sqrt{n}}$$

or $$t = \frac{\bar{X}_i - \mu}{s/\sqrt{n}}$$

As before we use the *population* standard deviation (σ) or the *sample* standard deviation (s) for calculating t, depending on the information we are given in the question.

t tables

The full ***t* tables** to be used in finding the appropriate critical values can be found in Appendix 5, p. 318.

It will be useful to go through our seven-step plan for a problem involving *small samples* and therefore the t distribution. Our steps are exactly as before, except in the way we calculate the critical values.

WORKED EXAMPLE **7.11**

A company requires batteries for its transport vehicles that last on average 20,000 kilometres before replacement. A battery company claims that it can supply batteries that meet this requirement. The buying company decides to test that claim and buys 16 batteries and finds that they last an average of 19,500 kilometres with a standard deviation of 1,200 kilometres. Test the suppliers claim against the buyers concern that batteries last less than 20,000 kilometres, at a 5 per cent level of significance.

Solution

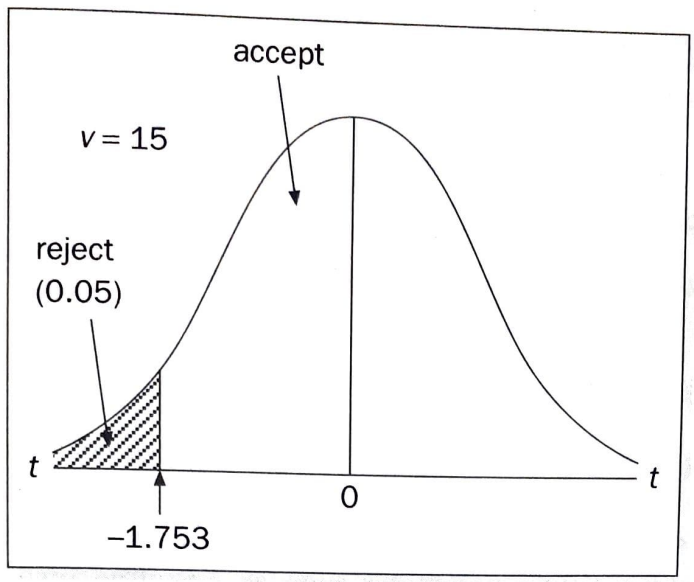
Note: This is arguably a *one-tail* test because the buyer is only worried if the batteries fall *below* the claim of 20,000 kilometres. The sample size is less than 32, so it is a t test example. We use s rather than σ in calculating standard errors.

Step 1 $H_0 : \mu = 20{,}000$ km
$H_1 : \mu < 20{,}000$ km

Step 2 $\alpha = 0.05$

Step 3 $\nu = n - 1$
$= 16 - 1$
$= 15$
$t_{0.05,15} = -1.753$

figure 7.15



Step 4 $t = \dfrac{\bar{X}_i - \mu}{\sigma/\sqrt{n}}$ (here s for σ)

$$= \frac{19{,}500 - 20{,}000}{1{,}200/\sqrt{16}}$$

$$= \frac{-500}{300} = -1.667$$

Step 5 $-1.667 > -1.753$

Step 6 Accept H_0

$\mu = 20{,}000$ km

Step 7 There is insufficient evidence for us to conclude that the batteries perform less well than is claimed.

SELF-CHECK QUESTIONS

7.9 It is believed that the mean time that third year students spend in the library per week is 18 hours. Assume that the time spent in the library is normally distributed. A random sample of 25 students gives a mean time of 18.6 hours with a standard deviation of 1.02 hours. Test the view that the mean time spent in the library is greater than 18 hours at the 5 per cent significance level.

7.10 Suppose we want to test, on the basis of a random sample of eight cartons, whether the average fat content of a certain make of ice cream is less than 15 per cent. The sample mean for fat content is found to be 14.3 per cent with a sample standard deviation of 0.96 per cent. What conclusion would you come to at the 1 per cent significance level?

7.11 Workers on a production line produce an average of 154 faulty items per shift. A training package has been developed and tested on a random sample of 20 workers. The twenty workers in the sample produce an

average of 141 faulty items with a standard deviation of 12 items. The training package will be applied to all workers only if it can be shown that it improves worker performance. Based on the sample should the firm proceed with the programme? Test at the 1 per cent level of significance.

Note: **Answers can be found on p. 351.**

Chi-squared test

The **Chi-squared (χ^2)** test looks not at an *individual* item of data (i.e. a single parameter) but at the *whole* distribution. As a result it is known as a *non-parametric* test, unlike the Z and t tests which are known as parametric tests with their focus on a single parameter such as a sample mean which is then compared with an allegedly known value of a parameter from the population.

For non-parametric tests, such as the Chi-squared test, the null hypothesis (H_0) and alternative hypothesis (H_1) are defined in terms of the distribution as a whole. Usually we test:

H_0 : the distribution as a whole is evenly (equally) spread.
H_1 : the distribution as a whole is *not* evenly (equally) spread.

χ^2 statistic

The **χ^2 statistic** or score makes use of actually *observed values* (O) which are then contrasted with the *expected values* (E) we should anticipate if the null hypothesis (H_0) is in fact true. It is defined as:

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O - E)^2}{E}\right]$$

where O = observed values
E = expected values (if null hypothesis, H_0, is true)
n = number of observations in the sample

The χ^2 distribution follows the pattern shown in Figure 7.16.

Note that, as with the t distribution, the χ^2 distribution actually changes as the sample size (n) changes. We therefore again make use of the idea of *degrees of freedom* (v) (see p. 170), with the number of degrees of freedom expressed as $v = n - 1$, where n corresponds to the number of observations in the sample.

Of course if the distribution as a whole *is* evenly spread throughout, as our null hypothesis H_0 suggests, then the *observed values* (O) would exactly equal the *expected values* (E). It follows that $(O - E)$ in the numerator of the χ^2 statistic will equal zero for every one of the n observations in the sample, and summing the squares of zero will still give us zero for χ^2. In other words χ^2 = zero will correspond to a *perfectly evenly spread* distribution. The less evenly spread the distribution, the greater the discrepancy between any observation (O) and its expected value (E), when the latter is calculated on the assumption that the distribution *is* evenly spread (i.e. H_0 is true).

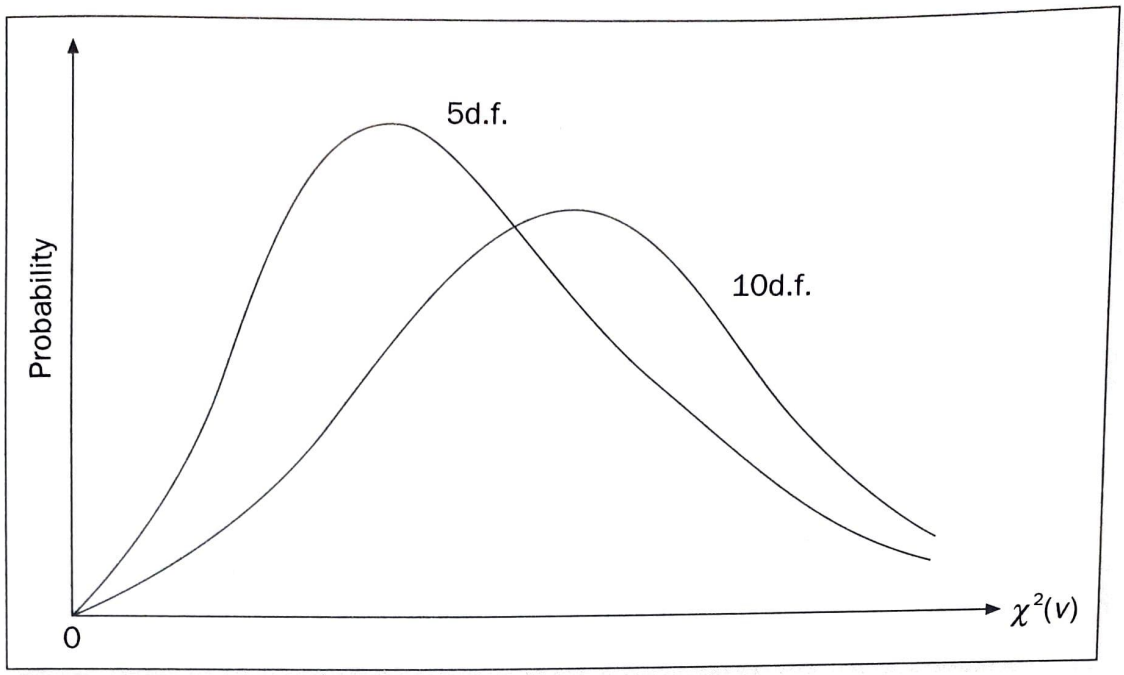


figure 7.16
χ^2 varies with v, where v is the number of degrees of freedom

Critical values

The **critical values** are those values for χ^2 outside of which we are willing to *reject* the null hypothesis (H_0) that the distribution is evenly spread. In this case we reject H_0, where the observations (O) are sufficiently different from the expected values (E) we should anticipate if the distribution really is evenly spread.

- As we have already noted the χ^2 distribution will vary with the *number of degrees of freedom* (v), where $v = n - 1$. It follows that the critical values will therefore vary with v.
- The *level of significance* (α), as before, represents the maximum risk we are willing to take in rejecting the null hypothesis (H_0) when it is in fact true.

The critical values for the χ^2 distribution can therefore be expressed as:

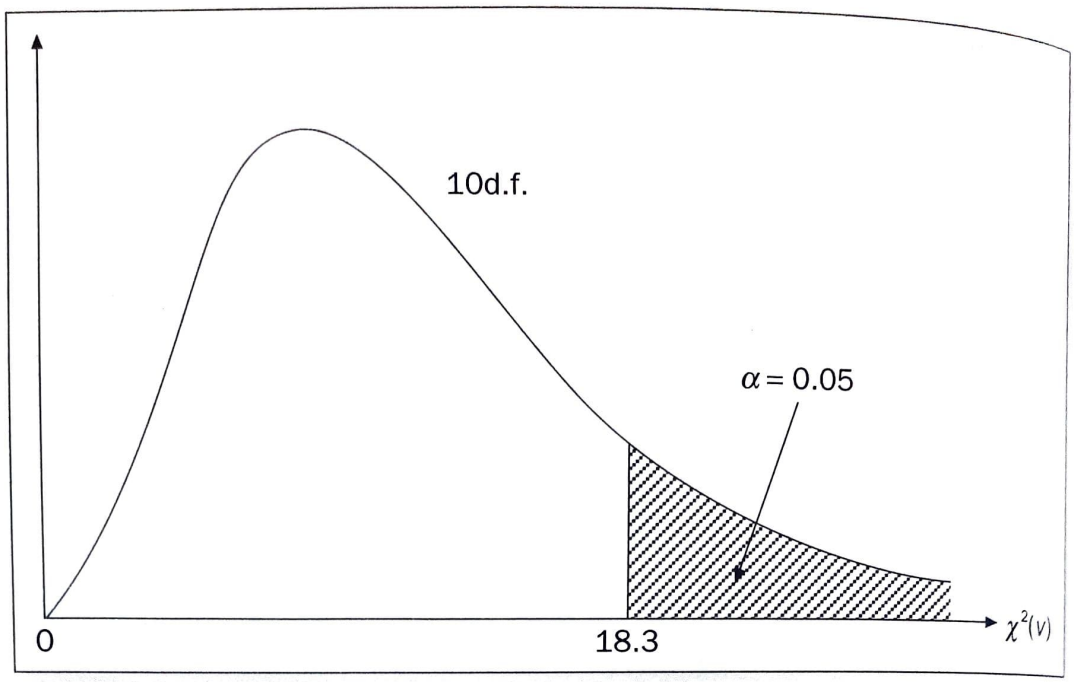
$$\chi^2_{\alpha,v}$$

where α = level of significance
v = number of degrees of freedom ($n - 1$)

All tests involving the χ^2 distribution are *one-tailed* tests.

Appendix 6 (p. 319) gives a full table of critical values for a given level of significance (α) and a given number of degrees of freedom (v). For example, as shown in Figure 7.17, for a 5 per cent level of significance ($\alpha = 0.05$) and 10 degrees of freedom ($v = 10$), we have a critical value of 18.3 for χ^2. In other words if, from our sample data, we calculate the χ^2 statistic as over 18.3, then we will reject the null hypothesis H_0 that the data is evenly spread. There will then be less than a 5 per cent chance of being wrong in deciding to reject the null hypothesis.

figure 7.17
finding the critical value of χ^2



WORKED EXAMPLE **7.12**

Over a year a large company experiences 100 computer breakdowns. These are classified into the four time-period slots indicated below.

Time of computer breakdowns	*Number of breakdowns*
9.00 a.m.–11.00 a.m.	22
11.00 a.m.–1.00 p.m.	18
1.00 p.m.–3.00 p.m.	26
3.00 p.m.–5.00 p.m.	34

Test the hypothesis that computer breakdowns are spread evenly through the working day, using a 5 per cent (0.05) level of significance.

Solution

We can follow the same seven steps in testing a hypothesis as for the Z and t distributions.

Step 1 H_0: Breakdowns evenly spread
H_1: Breakdowns not evenly spread

Step 2 $\alpha = 0.05$

Step 3
$$v = n - 1$$
$$= 4 - 1$$
$$= 3$$
$$\chi^2_{0.05,3} = 7.81$$

Step 4 Calculate the χ^2 statistic from the sample data

$$\chi^2 = \sum_{i=1}^{4} \frac{(O - E)^2}{E}$$

O	E	$(O - E)$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
22	25	−3	9	0.36
18	25	−7	49	1.96
26	25	+1	1	0.04
34	25	+9	81	3.24
				5.60

Note: If H_0 is true, then we expect 25 breakdowns to occur in each of the four time-periods of two hours.

$$\chi^2 = 5.60$$

Step 5 $5.60 < 7.81$

Step 6 Accept H_0

Step 7 Cannot reject H_0 that computer breakdowns are evenly spread over time. In other words there is not enough evidence to conclude that the computer breakdowns that occur at different times are due to anything other than chance.

Data shown in table or matrix form

Often the data are shown, or can be represented, in table or matrix form. As we shall see from Worked Example 7.13 we follow exactly the same procedures as above, with one difference which relates to calculating the *number of degrees of freedom* (v):

$$v = (r - 1) \times (c - 1)$$

where v = number of degrees of freedom
r = number of rows in table
c = number of columns in table.

As we shall also see in Worked Example 7.13, we can devise a useful formula for calculating the *expected values* (E) when data are presented in tabular form:

$$E = \text{row total} \times \frac{\text{column total}}{\text{grand total}}$$

where E = expected value

WORKED EXAMPLE **7.13**

A survey of job satisfaction was undertaken for systems analysts and computer programmers employed by a large firm. A questionnaire asked 100 workers of each type whether or not they were satisfied with their job. The results are shown below:

Type of employee	*Satisfied*	*Not satisfied*	*Totals*
Systems analysts	63	37	100
Computer programmers	53	47	100
Totals	116	84	200

Is there a significant relationship between the type of employee and job satisfaction? Test at a 5 per cent level of significance.

Solution

This is a χ^2 test since we are testing the null hypothesis (H_0) of no association between type of employee and job satisfaction against the alternative hypothesis (H_1) that there is such an association. We are using *observations* from sample data (200 employees) to help test H_0.

If 'everything is in proportion' and there is no association, then H_0 is true. We should then *expect* that the *overall proportion* of satisfied employees should be reflected in the totals for each type of employee claiming to be satisfied. We return to this idea below when working out the 'expected' values (E).

Sometimes tabular data are given with row and column totals provided. If not, it will help if you insert such row and column totals. Note also that the sum of the respective row and column totals is referred to as the 'grand total', here the 200 employees involved in the survey.

We can follow our usual seven steps, but remembering to apply the procedures already mentioned for data in tabular form:

Step 1 H_0: no association between type of employees and job satisfaction (i.e. responses are in proportion).
H_1: association between type of employee and job satisfaction (i.e. responses are not in proportion).

Step 2 $\alpha = 0.05$

Step 3 $v = (r - 1) \times (c - 1)$
$v = (2 - 1) \times (2 - 1)$
$v = 1$

(*Note:* when using tabular data, do *not* count the row total or column total when calculating v)

$\chi^2_{\alpha,v} = \chi^2_{0.05,1} = 3.84$

Step 4 Calculate the χ^2 statistic from the sample data.

$$\chi^2 = \sum_{i=1}^{4} \frac{(O - E)^2}{E}$$

where O = observations
E = expected values if H_0 (no association) is true.

Here we can apply our formula to work out expected values.

$$E = \text{row total} \times \frac{\text{column total}}{\text{grand total}}$$

$$E \text{ (Systems analysts satisfied)} = 100 \times \frac{116}{200} = 58$$

$$E \text{ (Systems analysts not satisfied)} = 100 \times \frac{84}{200} = 42$$

$$E \text{ (Computer programmers satisfied)} = 100 \times \frac{116}{200} = 58$$

$$E \text{ (Computer programmers not satisfied)} = 100 \times \frac{84}{200} = 42$$

The formula simply helps us impose the *overall proportions* of satisfied/not satisfied on the totals for systems analysts and computer programmers respectively. The resulting values tell us what to expect if the null hypothesis H_0 (no association) is true.

We can express our table of expected values (E) as:

Expected values

	Satisfied	Not satisfied
Systems analysts	58	42
Computer programmers	58	42

We can now calculate χ^2, as follows:

O	E	$(O - E)$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
63	58	+5	25	25/58 = 0.431
37	42	−5	25	25/42 = 0.596
53	58	−5	25	25/58 = 0.431
47	42	+5	25	25/42 = 0.596
				2.054

$\chi^2 = 2.054$

Step 5 $2.054 < 3.84$

Step 6 Accept H_0

Step 7 No association between type of employee and job satisfaction. There is not enough evidence to conclude that these different types of employees have different views as to job satisfaction.

Note: Review question 7.15 is a χ^2 question.

REVIEW QUESTIONS

7.1 The manageress of a restaurant claims that her weekly turnover is £3,130 with a standard deviation of £115 per week. If the turnover is normally distributed and a random sample of 25 weekly turnover figures is recorded, what is the probability that the sample mean will be:

(a) less than £3,100? (b) greater than £3,190? (c) between £3,110 and £3,180?

7.2 The mean salary of business studies graduates, one year after graduating, is £12,500 with a standard deviation of £1,300. What is the probability that a random sample of 49 of these had a mean salary of at least £13,000?

7.3 A sample of 49 observations is taken from a normal population. The sample mean is 58, and the standard deviation of the sample is 10. Determine the 95 per cent and the 99 per cent confidence intervals for the population mean.

7.4 A wine shop owner wishes to estimate the mean number of bottles sold to account-holders per month. A sample of 60 accounts gives a mean of 8.6 bottles per month with a standard deviation of 2.3. Develop a 99 per cent confidence interval for the population mean.

7.5 The Durable Construction Company produces frames for solar panels. The frames need to withstand certain pressures. A sample of 45 frames is taken from the production line and tested for the weight they can support before they distort. The test results from the sample of 45 frames suggest an average weight of 39.8 kilos before distortion with a standard deviation of 6 kilos.

(a) Construct a 99 per cent confidence interval for the population mean.
(b) Construct a 90 per cent confidence interval for the population mean.
(c) How large should the sample be if we want a 99 per cent confidence interval for the population mean with a margin of error of 2 kilos?

7.6 Transport economist, Sisson, wishes to estimate how often residents of Cambridge use the city bus system per month.

(a) How many residents must be surveyed if he wants to develop a 95 per cent confidence interval with a margin of error of three trips per month? A previous pilot study gave the standard deviation as 14 trips.
(b) If Sisson randomly selects 160 residents and obtains a mean of 22.4 trips with a standard deviation of 15, construct a 95 per cent confidence interval for the population mean.

7.7 The owner of a restaurant believes that the average amount spent on a meal per couple is £50.00. She takes a random sample of 40 couples and finds that the mean expenditure is £53.70 with a standard deviation of £7.80. Test the owner's claim at the 5 per cent level of significance.

7.8 A car manufacturer states that its new model does 32.5 mpg. A recent independent study of 50 of the new cars selected randomly showed a mean mpg of only 30.4 with a standard deviation of 5.3 mpg. Test the view that the manufacturer's claim is too high using a 1 per cent level of significance.

7.9 The expected life of a car engine is 90,000 miles. A random sample of 190 cars gives a mean life of 96,700 miles with a standard deviation of 37,500 miles. Test whether the cars are lasting longer than expected at the 1 per cent level of significance.

7.10 An estate agent advertises that it can sell houses in 40 days or less. A random sample of 64 recently sold homes shows a sample mean selling time of 45 days with a standard deviation of 20 days. Using the 5 per cent level of significance, test the validity of the agent's claim against a rival's claim that this is untrue.

7.11 A poll of 121 randomly selected students reveals that the mean number of CDs they buy a year is 7.01 with a standard deviation of 3.74. Test the claim that the mean for all students is less than 7.5 CDs at the 1 per cent level of significance.

7.12 A machine is set to cut metal rods to a length of 8.75 cms. It is felt by the operator that the rods are too short. The hypothesis is to be tested at the 1 per cent significance level from the following sample of rods: 9.0, 8.5, 8.7, 8.4, 8.6, 8.3, 8.5 and 8.8. What conclusion can be drawn on the basis of the sample?

7.13 A mail order firm claims that new procedures enable it to dispatch orders more quickly than previously. Records show that the previous average delay was 20 days. A random sample of 17 orders using the new procedure showed a delay time of 18 days with a sample standard deviation of 2.5 days. Is the claim made for the new procedures supported at the 5 per cent level?

7.14 A recruitment agency claims that four bedroom detached houses near the Cambridge Science Park can be rented for £900 per month on average. A random sample of 20 houses gave a mean rental of £1,050 per month for this type of property with a sample standard deviation of £290 per month. Test the agency's claim against the claim that rentals are higher at the 10 per cent significance level.

7.15 Using the following observations recorded by a researcher, check if the assumption of no relation between lung cancer and smoking can be accepted at the 5 per cent level of significance.

Observation table

	Smokers	*Non-smokers*	*Totals*
Cancer sufferers	200	200	400
Non sufferers	150	450	600
Totals	350	650	1,000

Answers to Review questions can be found on p. 425.

Further study and data

Texts

Bancroft, G. and O'Sullivan, G. (1993), *Quantitative methods for accounting and business studies*, 3rd edn, McGraw Hill, chapters 15–18.

Curwin, J. and Slater, R. (1996), *Quantitative methods for business decisions*, 4th edn, International Thompson Business Press, chapters 13–16.